



Air Quality Data Management and Integration System Scoping Study

Report to Defra and the Devolved Administrations

Unrestricted
ED46602
Issue 1.2 AEAT/ENV/R/3005
August 2010


Title	Air Quality Data Management and Integration System - Scoping Study
Customer	Defra and the Devolved Administrations
Customer reference	RMP 5603
Confidentiality, copyright and reproduction	This report is the Copyright of Defra and the Devolved Administrations and has been prepared by AEA Technology plc under contract to Defra and the Devolved Administrations dated January 2010. The contents of this report may not be reproduced in whole or in part, nor passed to any organisation or person without the specific prior written permission of Defra and the Devolved Administrations. AEA Technology plc accepts no liability whatsoever to any third party for any loss or damage arising from any interpretation or use of the information contained in this report, or reliance on any views expressed therein.
File reference	
Reference number	ED46602- Issue 1.2

AEA group
The Gemini Building
Fermi Avenue
Harwell
Didcot
Oxfordshire
OX11 0QR

t: 0870 190 6602
f: 0870 190 6377

AEA is a business name of AEA Technology plc

AEA is certificated to ISO9001 and ISO14001

Author	Name	Andrew Monteith, Ollie Cronk, Rachel Yardley, Paul Willis, Xingyu Xiao
Approved by	Name	Paul Willis
	Signature	
	Date	August 2010

Executive summary

Fewer than half of the UK's air quality dataset users have access to all of the air quality monitoring, modelling and emissions data that they require, and the majority do not have access to the associated information which is needed to provide context and relationships between the causes and impacts of air pollution, for instance, Met data, traffic and land use statistics, population and health data.

To overcome this barrier and to maximise the overall availability and use of the data it would be possible for the UK to integrate these datasets. This report summarises the findings of a scoping study undertaken in 2010 to investigate the feasibility of such an integration process, and makes recommendations on how this could be achieved.

Benefits of the proposed integration system

The proposed integrated system will increase data processing efficiency and reduce operating costs. Air quality and other related data will be more easily available and accessible to members of the public and the air quality community. Statutory data reporting procedures will be simplified. The amount of time spent searching for, manipulating and interpreting data will be greatly reduced. It will be possible to develop useful tools to aid policy-makers, making use of a wider pool of data than has previously been possible. Examples of data visualisation and analysis tools currently used in Europe are shown in Section 1.3.

The integration of the UK's air quality data will help to meet our regulatory obligations, by allowing and promoting the reuse of public data, simplifying reporting procedures and standardising data formats. The proposed system architecture is compliant with the requirements of INSPIRE and the UK Location Programme. The key regulatory drivers are discussed in Appendix 1.

Data considered in the study

This scoping study focuses on the top priority data sets of the National Atmospheric Emissions Inventory, Pollution Climate Mapping, Automatic Urban and Rural Network, London Air Quality Network and non-automatic network data (hydrocarbons, heavy metals, black smoke), and examines how maximum efficiency and benefit can be gained through the integration of these data.

Also key to the success of developing or improving the efficiency of air quality tools is the availability and accessibility of essential non-air quality data sets, in particular Met data and traffic data. Although Met data is currently not freely available, there are several publicly available traffic datasets, which could be made available in a straightforward manner to an integrated air quality platform, through liaison with the data providers. Both met and traffic data are priorities for integration.

Stakeholder feedback

Many stakeholders identified the same shortfalls and listed the same type of datasets that are essential to get the most value out of the UK's air quality data assets. The primary difficulty is knowing what data are available in the first place and then being able to find them. Better search and indexing tools are essential to allow users to see and search the datasets.

81% would welcome a single air quality portal with improvements to the availability, format and integrity of these data. **75%** would welcome basic or complex online tools to help analyse these data, however, in making decisions on air quality we would caution against the setting of an automated response, which undermines the technical excellence of air quality expertise in the UK.

During the development of the proposed data integration structure, further stakeholder engagement will be required during the early stages to ensure a high level of commitment and to ensure the end results is focused on user requirements.

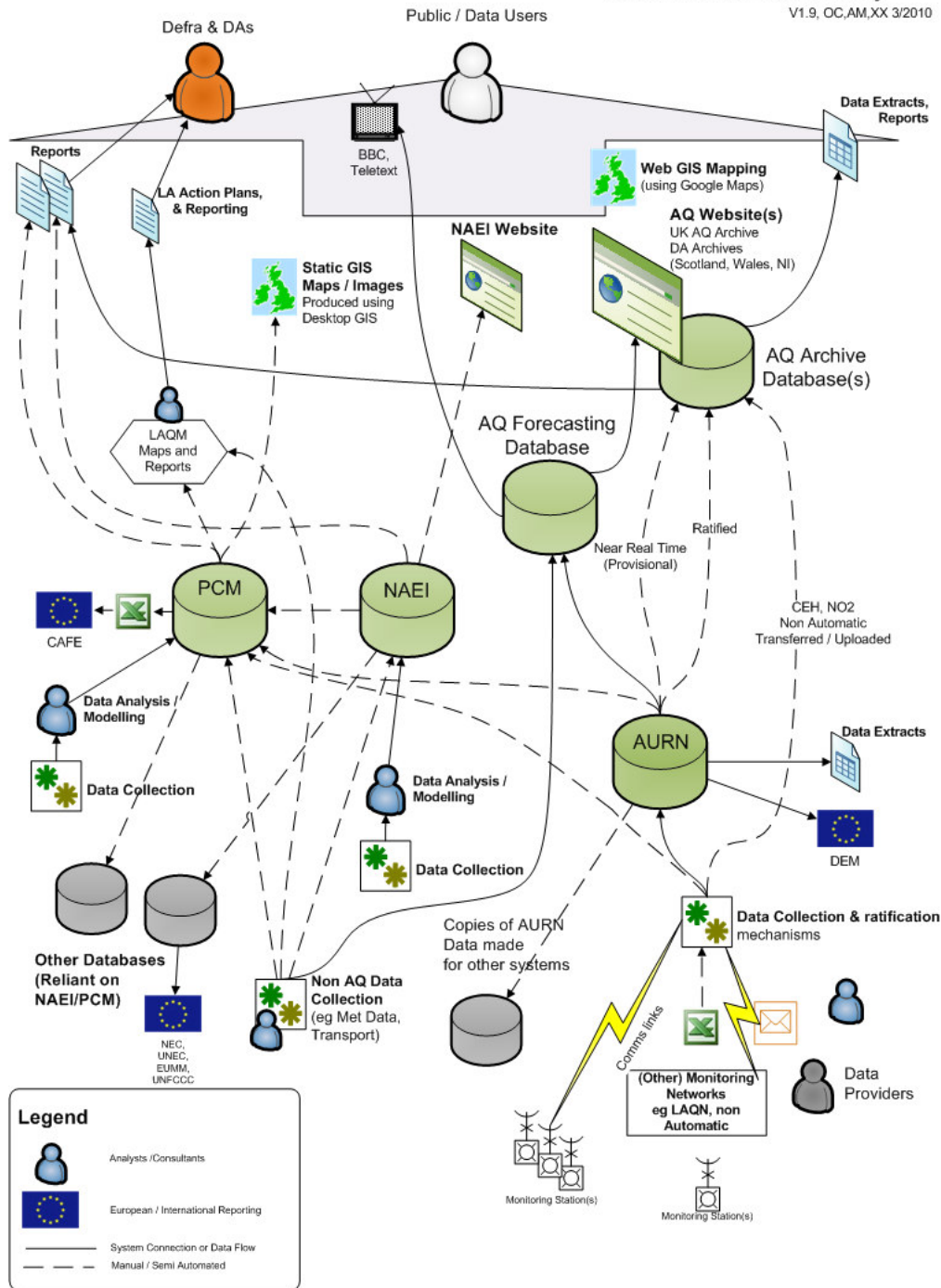
The current situation

Three key problems have been identified with the current situation. These are:

- The lack of structure in the architecture of the overall system of air quality and other datasets
- The disparity of the datasets in terms of standard formats
- Lack of or inconsistent metadata

Current AQ Summary

V1.9, OC,AM,XX 3/2010



Development of tools to make better use of existing data

Section 4 of this report describes the complexity of the current situation, the structure of the datasets, the currently available tools and their limitations. In an integrated system the data inputs to the current models and tools could be fully automated, and the resulting data, reports, plans and maps would automatically feed into Local Air Quality Management and other tools, saving processing time and reducing the element of human error.

One of the benefits of a fully integrated system for air quality data would be the easier application of online tools to view, sort and analyse the data to give them meaning and make them applicable to real life. Suggestions are given in Section 4.8 under the following categories:

- Local Air Quality Management or national air quality compliance tools
- Presentational tools
- Practical tools
- Action planning and impact analysis tools
- Emissions scenario testing tools
- Analysis tools

Proposed Approach

The proposed approach provides a platform for tool development to allow the air quality community and members of the public to make the best use of existing data. It also provides a future-proof solution for standardising and integration air quality datasets in a way that is compliant with European legislation and initiatives.

There are costs associated with:

- The creation of a detailed service architecture, development of data standards and other over-reaching activities which will impact all datasets, in the range of £60K to £100K depending on scope
- The creation of good metadata, which is vital to make the data useful, is likely to be a major cost, in the range of £120K to £200K. It is advised that this is undertaken before any modifications occur on individual datasets.
- The changes made to individual datasets. Costs per dataset have been illustrated in Section 5.1 as between £25K and £75K per dataset in addition to the costs that have occurred due to creation of good metadata.
- Implementation of the web services themselves will also be a significant cost. This is dependent on how far standard tools can be used to meet the specification, and the level of appropriate skills the data providers have to implement them.

Through the use of industry standard formats and definitions of metadata the proposed architecture will allow for the system to be modified to meet changing requirements rather than fixed systems designed around specific requirements from a set era. The proposed approach will have the following benefits:

- More automation and less manual intervention. Less time will be spent searching for data and manipulation of data will be quicker due to standardised formats and new tools
- Reduced operating costs
- Robust platform for developing new products and services to add value to existing data
- Simpler and more automated reporting lines
- Ability to link or overlay data from different datasets
- Better, informed decision-making, using a wide range of datasets

Quick Wins

It would be possible to start to add value to the data and make them more useable in the short term, without the bottom-up standardisation approach which has been outlined above. A number of quick wins have been identified by this scoping study and are described in Section 6.4, this is a different approach to the proposed approach and the costs as such reflect different work that will be undertaken. In some occasions the work will overlap with the approach stated above thus some of the costs defined below could be incorporated into the costs defined above:

- Using this scoping study as a starting point, catalogue all useful air quality and non-air quality datasets and metadata, including those, but not exclusively those owned by Defra and the Devolved Administrations. Create appropriate web pages to list and provide links to these datasets. We anticipate the cost for this to be in the region of £25k.
- Integration of the continuous air quality data is a prime candidate for early integration. The data is relatively simple, the measurements are important and we have a lot of them. One option for this integration is to make use of RDFa – this is where RDF (a variant of XML) is added to existing HTML pages designed for human consumption (invisibly to human users) that adds support for use of the data systematically. We anticipate the cost for this to be in the region of £50k per dataset.
- Some data providers (KCL and BADC) hold local Met data. This could also be integrated with relatively little effort but with a great benefit (access to Met data has been identified by data users as a high priority). In particular, KCL have about 20 met sites in London and the South East whose data could be ready for integration in the short term. We anticipate the cost for this to be less than £100k.
- The collection and integration of metadata is viewed as a very high priority. Making these data available and easily accessible online is the first step to provide more meaning to the data. Standardising the format of this metadata and ensuring its completeness is important, but a more difficult and longer term task. We anticipate the cost for this to be in the region of £120k.
- If DfT can provide transport data in a consistent format these could be use to plug into LAQM tools. Further investigation is required to determine how DfT data could be readily transformed and automatically applied to currently available tools, but this is not expected to be a difficult nor expensive task. We anticipate the cost for this to be in the region of £25k.
- Identify reporting processes which can be readily simplified and automated and develop simple tools to do this.

Proposed Future AQ Architecture Outline

V1.9, OC,AM,XX 3/2010

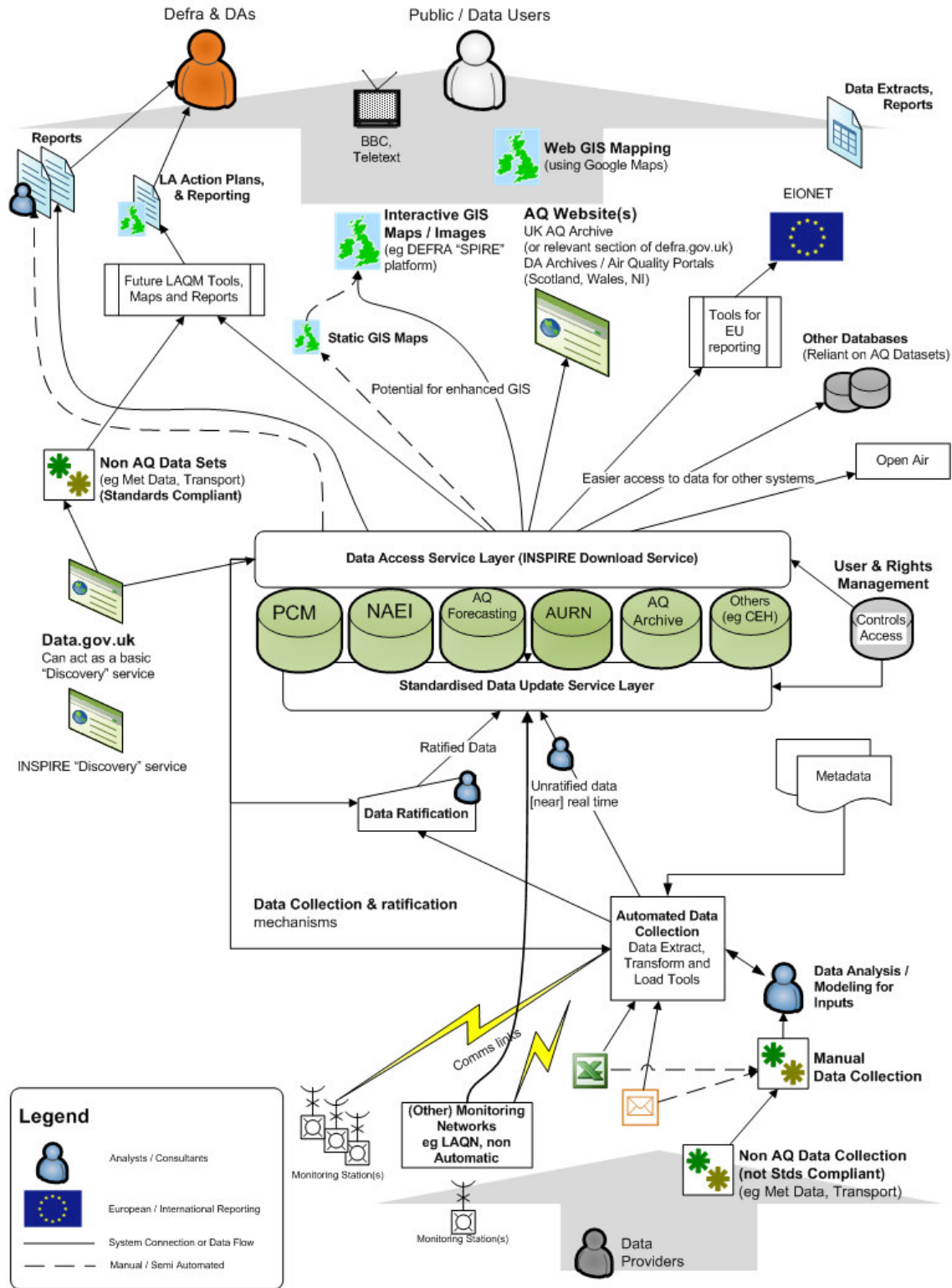


Table of contents

1	Introduction	11
1.1	Drivers	12
1.2	Objectives	13
1.3	Existing air quality data toolsets	14
2	Methodology	17
2.1	Workshops and data gathering	17
2.2	Steering Group	18
2.3	Acknowledgements	19
3	Data to be integrated	20
3.1	Top priority data	20
3.2	Mid priority data	22
3.3	Low priority	22
3.4	Non air quality data	22
4	Current Situation	26
4.1	Overview	26
4.2	Overview Diagram	27
4.3	Assessment of Air Quality Datasets	28
4.4	Current Architecture for Tools	30
4.5	Local Air Quality Management Tools	31
4.6	Tools for Public Information and Research	34
4.7	National Air Quality Assessment Tools	37
4.8	Stakeholder Survey	39
5	Proposed Future Approach & Architecture	44
5.1	Changes Required	44
5.2	Underlying Principles	45
5.3	Overview of proposed future Architecture	48
5.4	Platform for future tool development	50
5.5	Challenges	51
5.6	Implications for stakeholders	51
6	Roadmap and Proposed Transition	53
6.1	Proposed Strategy / Implementation Tasks	53
6.2	Costs	55
6.3	Gantt Chart	55
6.4	Quick Wins	55
6.5	Alternative Strategy	56

Appendices

Appendix 1	Regulatory Drivers
Appendix 2	Workshop 1 Minutes
Appendix 3	Data User Survey
Appendix 4	Data Provider Survey
Appendix 5	Workshop 2 Minutes
Appendix 6	Dataset Summary and Scoring Criteria
Appendix 7	INSPIRE, SEIS and data.gov.uk

1 Introduction

Defra and the Devolved Administrations own a large amount of historic and current data on air quality and other environmental factors, which are managed by a small number of contractors. Each contractor, and in some cases each dataset, has a different system to capture and manage the data, which makes data analysis, reporting and decision making difficult.

The majority of these data are produced using public funds, yet many are not readily accessible to the general public.

Fewer than half of the UK's air quality dataset users have access to all of the air quality monitoring, modelling and emissions data that they require, and the majority do not have access to the associated information which is needed to provide context and relationships between the causes and impacts of air pollution, for instance, Met data, traffic and land use statistics, population and health data.

To overcome this barrier and to maximise the overall availability and use of the data it would be possible for the UK to integrate these datasets. This report summarises the findings of a scoping study undertaken in 2010 to investigate the feasibility of such an integration process, and makes recommendations on how this could be achieved.

The aims of the proposed data integration process are to:

- Maximise the overall availability and use of the data to support stakeholder objectives
- Standardise definitions and data formats
- Standardise data updates
- Catalogue datasets and disseminate information about the data that are available
- Allow different datasets and different systems to be interrogated as one standard system
- Ensure that data produced with public money are made available to the public.

The proposed integrated system will increase data processing efficiency and reduce operating costs. Air quality and other related data will be more easily available and accessible to members of the public and the air quality community. Statutory data reporting procedures will be simplified. The amount of time spent searching for, manipulating and interpreting data will be greatly reduced. It will be possible to develop useful tools to aid policy-makers, making use of a wider pool of data than has previously been possible.

We propose that a new data approach and architecture that incorporates a spatial data infrastructure are developed to enable access to the UK air quality data assets that are held by several contractors. This data infrastructure will allow the current disparate geographic information systems better align and integrate. The proposed infrastructure will fully comply with requirements laid out in the INSPIRE Directive¹, thus making the UK air quality data compatible with all other spatial data in the UK.

This report describes the current situation in the UK, with many data frameworks, data flows and architectures. It investigates the issues associated with these disparate datasets and identifies common difficulties encountered by researchers, consultants and other groups who need to access and use the data. From this the report lays out a vision for the future, considering options for the integration of the UK's air quality data and recommending a solution. We outline the key considerations and steps towards achieving this ultimate integrated system and discuss the possible options for tool development based on the integrated data.

Where possible the report indicates approximate costs and timescales for the delivery of the integrated system.

¹ Directive 2007/2/EC of the European Parliament and of the Council

1.1 Drivers

There are many reasons why Defra and the Devolved Administrations are considering the integration of all UK air quality data. Broadly these can be sorted into two groups:

1. The potential benefits for all stakeholders, including Government, the air quality community and members of the public
2. The regulatory requirements to standardise datasets and maximise availability and reuse of publicly funded data

1.1.1 Benefits

The benefits of integrating the UK's air quality data assets are many, some of which are given in Table 1.1.

Table 1.1 Potential benefits of data integration

Monetary
<ul style="list-style-type: none"> ➤ Increased efficiency of data providers and data users due to more automation and less manual intervention. Less time will be spent searching for data ➤ Quicker manipulation of data due to standardised formats and new tools ➤ Reduced operating costs for data providers, data users and Defra and the Devolved Administrations ➤ Potential for developing new products and services to add value to existing data ➤ Growth and opportunities in the market for tools and services ➤ A more cost effective and efficient approach to reporting at an International level (e.g. obligations under EC directives) using automated procedures
Process and quality improvements
<ul style="list-style-type: none"> ➤ Increased collaboration between stakeholders ➤ Modernised workflows and future-proofing our data management system, The proposed integrated system will be INSPIRE compliant and use the latest technology ➤ Quicker and easier analysis, linking or overlaying data from different datasets so the correct conclusions can be drawn more easily from the data, and more factors can be considered ➤ Better, informed decision-making, using a wide range of datasets
Public
<ul style="list-style-type: none"> ➤ Provision of access to data and reduced restrictions ➤ Improved visualisation tools could be developed from the proposed platform ➤ Increased transparency of UK Government ➤ More effective emergency response as a result of data being more readily available ➤ Possible raised public awareness and understanding of air quality issues ➤ Possible improvements in public health

1.1.2 Local Air Quality Management Review

Defra's In House Policy Consultancy (IHPC) has recently carried out a review of Local Air Quality Management procedures². The final report is expected to highlight the following:

- The need to improve links to other policy areas including health, transport, land-use planning and climate change, which would become more feasible with the integration of air quality and non air quality datasets
- The need to streamline the Review and Assessment process and reduce the labour costs involved
- The need to integrate local and national information and create a larger national pool of monitoring data
- A recommendation for Government to routinely publish a fuller statistical overview of ambient air quality trends to match the published trend information on emissions, combined with better information on the health impacts. With an integrated system and associated toolset the statistical analyses could be generated quickly, easily and automatically, as frequently as is required.

These recommendations would all be addressed by the implementation of the proposed data integration system.

1.1.3 Regulatory Requirements

The key regulatory drivers of this scoping study and subsequent implementation of an integrated system for air quality and other environmental data, are listed below and described in more detail in Appendix 1:

- EU Directive on the Reuse of Public Sector Information
- CAFE Directive
- INSPIRE Directive
- UK Location Strategy
- Shared Environmental Information System

The integration of the UK's air quality data will help to meet our regulatory obligations, by allowing and promoting the reuse of public data, simplifying reporting procedures and standardising data formats. The proposed system architecture is compliant with the requirements of INSPIRE and the UK Location Programme.

1.2 Objectives

The key objective of this scoping study is to investigate and define what needs to be done to create a data infrastructure that ensures the accessibility and re-usability of Defra's and the Devolved Administrations' key air quality data investments.

It has been agreed with Defra and the Devolved Administrations that any data infrastructure should:

- Be compliant with the initiatives of the Defra INSPIRE Team
- Be generally "future proof" by coordinating with any other required future data harmonisation plans
- Set out a pathway for other data sets to be incorporated into, such as those outside of Air Quality
- Readily permit inter-comparison of initiatives and data generated at local, national and EU level
- Permit analysis of the impact / potential of measures.

Specific objectives of this scoping study are to:

- Examine data presentation issues, in respect of geographic presentation and online tools to analyse the results
- Discuss how it might be possible to move towards a common Data infrastructure for all Defra contractors

² Review of Local Air Quality Management: final report. January 2010, IHPC

- Consider options for tools to allow presentation and integration of past and current data, and projections where available, and other optional reporting or dashboard type tools including action planning and cost effectiveness tools
- Identify and map key dependencies between the data sets and possible desired outcomes
- Outline a new comprehensive data model that is compliant with the Defra INSPIRE plans, and a path towards that model
- Analyse the cost of implementation of the new data model and features, the ease of implementation of each measure, and risk of failure
- Lay out options for possible next steps.

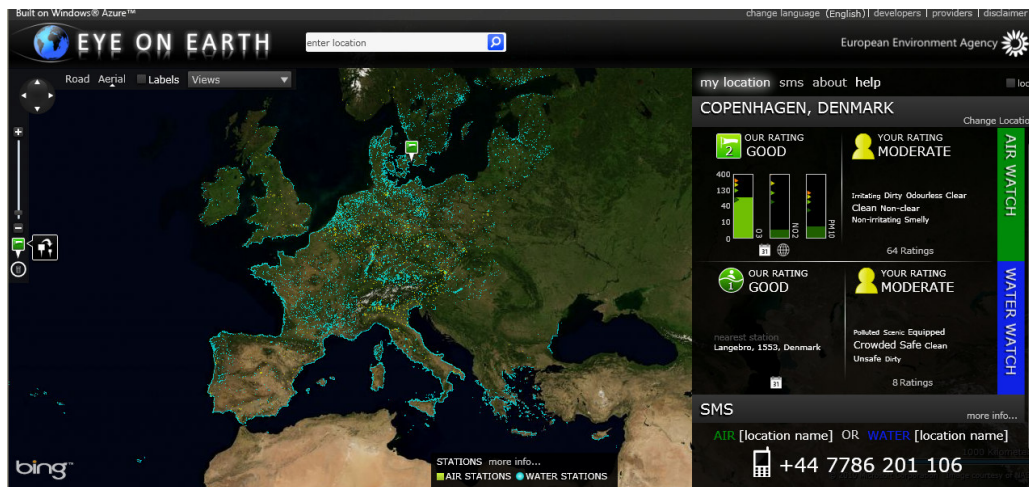
1.3 Existing air quality data toolsets

This section highlights some high profile tools that have been developed for air quality data around the world, thus serving to demonstrate what could be achieved if the UK datasets were integrated, with standardised formats and architecture.

Eye on Earth

The European Environment Agency's Eye on Earth is a communication tool that presents real time scientific water and air quality data alongside the observations of members of the public on an interactive map using a Microsoft platform. The tool has been developed so that, over time, additional datasets may be added to turn the Eye on Earth into a 'global observatory for environmental change'. It incorporates data from across Europe, from multiple data providers.

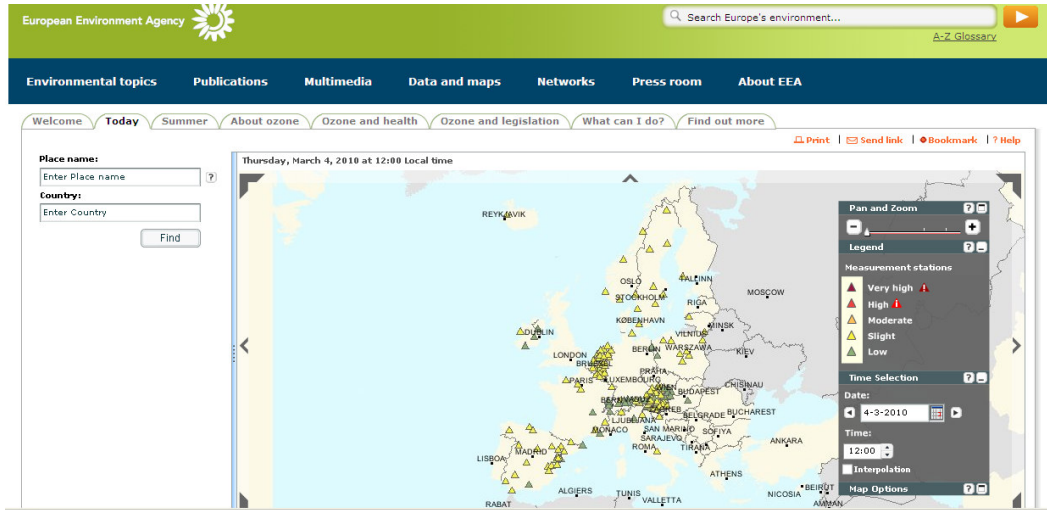
Figure 1.1 Eye on Earth screenshot (<http://eyeonearth.cloudapp.net/>) 04/03/2010



Ozoneweb

The European Environment Agency's Ozoneweb provides the public with easy access to information about ground level ozone pollution across Europe via an interactive map and graphical tools. The information is based on near real-time data provided by national and regional organisations, whose datasets are compatible and are able to be interrogated by Ozoneweb software.

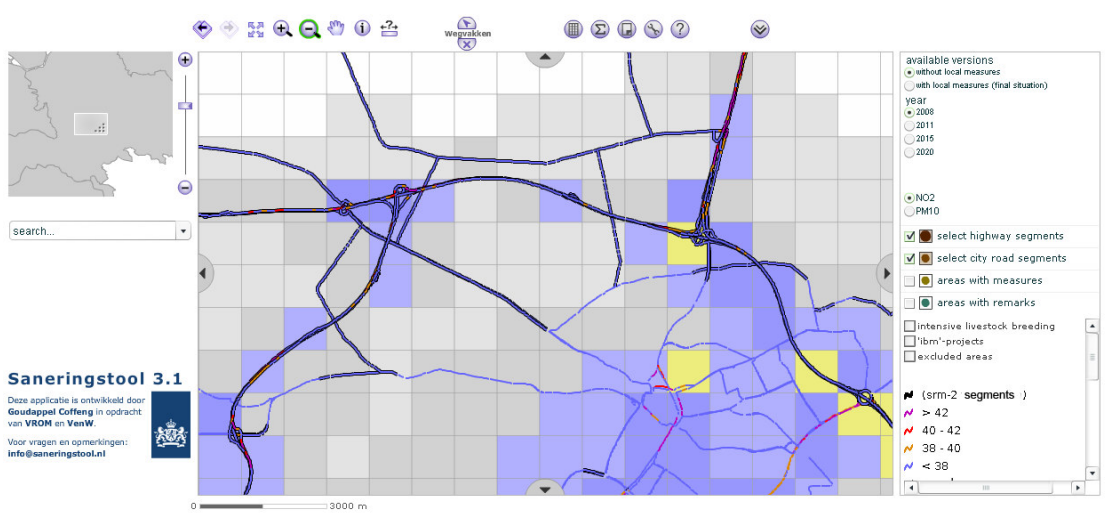
Figure 1.2 Ozoneweb screenshot (<http://www.eea.europa.eu/maps/ozone/welcome>) 04/03/2010



Saneringstool

The Dutch Ministry of Housing, Spatial Planning and Environment's Clean Air Policy Tool maps NO₂ and PM₁₀ concentrations along the entire Dutch road network. The tool allows policy makers to investigate the impact at any site along the roads of implementing regional and location-specific mitigation measures. The toolset includes analysis of the current data and scenario testing for the next ten years. This has been achieved through a co-operation programme between central government and the provincial and municipal authorities.

Figure 1.3 Saneringstool screenshot (http://www.saneringstool.nl/saneringstool_ENG.html) 04/03/2010



INSPIRE Geoportal

The European Commission's INSPIRE Geoportal allows users to search for spatial data sets and spatial data services, and subject to access restrictions, view and download spatial data sets from the EU Member States. The available spatial data comes from a wide range of sectors governed by INSPIRE.

Figure 1.4 INSPIRE Geoportal screenshot (<http://www.inspire-geoportal.eu/>), 05/03/2010

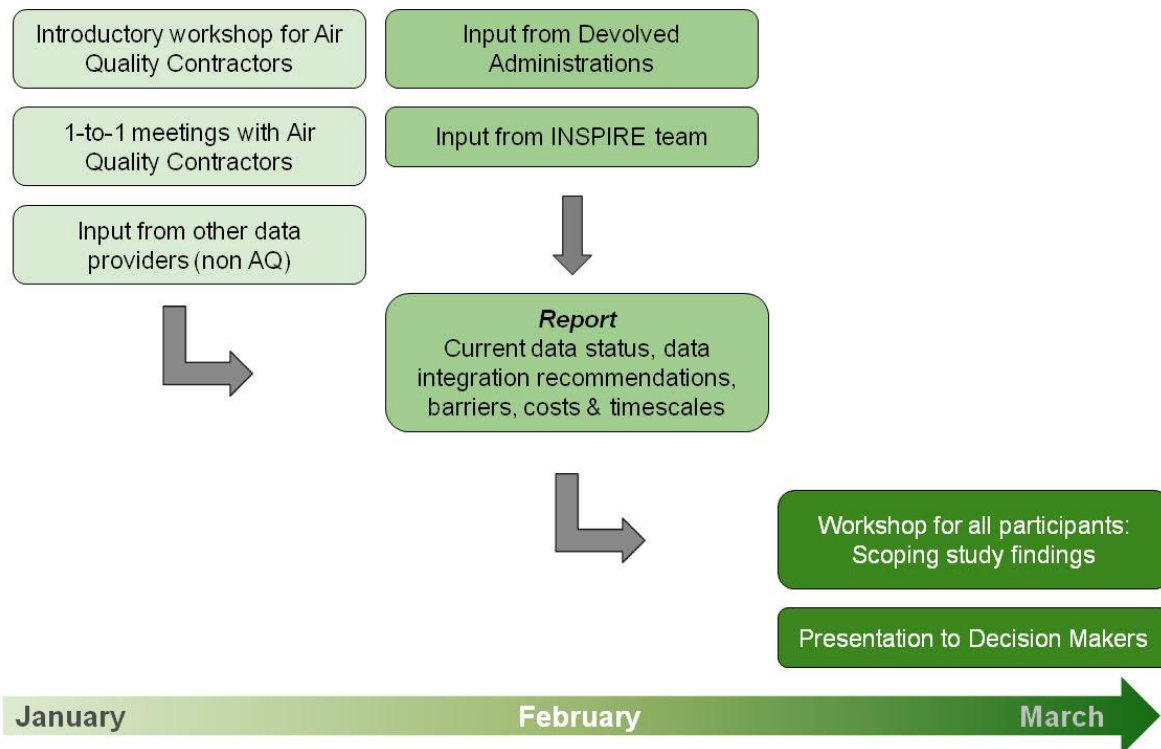


2 Methodology

2.1 Workshops and data gathering

The UK air quality data assets are produced and managed by several private contractors, who will play a key role in any future data integration exercise. Currently each contractor manages a number of datasets, and between contractors there is no requirement for a consistent approach. In order to engage with the contractors and to gather the necessary information to conduct this study, two stakeholder workshops and other stakeholder involvement events were held between January and March 2010.

Figure 2.1 Overview of scoping study project flow chart



2.1.1 Introductory Workshop

January 2010

An introductory workshop was held at Defra, Ergon House, for the contractors managing high priority datasets. The purpose of the workshop was to:

- Introduce the concept of the scoping study
- Gain support and engage with the data providers
- Ensure a common understanding of the aims of the scoping study and the case for data integration
- Present and discuss the current datasets, data management practices and barriers to integration
- Discuss common problems with availability and analysis of air quality data
- Brainstorm potential user tools which may be developed for use with an integrated dataset.

Minutes of the January 2010 workshop are attached in Appendix 2.

2.1.2 Briefing with Defra and the Devolved Administrations

February 2010

A briefing was held at Defra on the 10th February. The purpose of the meeting was to:

- Introduce the scoping study and progress to date to the Devolved Administrations and the Defra INSPIRE team
- Gain feedback on the study objectives and direction from the Devolved Administrations
- Ensure that the objectives of the study and the proposed data infrastructure align with the INSPIRE Directive requirements and its implementation within the UK
- Get input and an update from the INSPIRE team on the progress of the INSPIRE implementation within the UK

2.1.3 Stakeholder Questionnaires

February 2010

It is very important that the data infrastructure specification and design are driven by the individuals and groups who use the data. Two questionnaires were released to gather more detailed information on the air quality data management systems currently used in the UK by Defra's and the Devolved Administrations' contractors, and the specific issues encountered by and requirements of the UK air quality data users. The Data Providers questionnaire was distributed to contractors involved with the production and management of the UK's high priority datasets, as defined in Section 4. The Data Users questionnaire was distributed to a selection of Local Authorities, air quality consultants and researchers. Both questionnaires are attached in Appendices 3 and 4.

2.1.4 Stakeholder Review of Draft Report

March 2010

The draft recommendations of this study have been reviewed by all stakeholders who participated, in particular the following groups:

Defra and the Devolved Administrations
Data Integration Scoping Study Steering Committee
High priority data providers

2.1.5 Final Workshop

March 2010

The final workshop was held at Defra in March 2010 to present the findings of the scoping study to the key stakeholders, to outline the next steps and to gain ongoing understanding and support for any future integration project.

Minutes of the March 2010 workshop are attached in Appendix 5.

2.2 Steering Group

An Ad-hoc steering group was formed to guide the study and review the project outcomes. Three experts in air quality were invited and agreed to participate, alongside representatives from Defra's Atmospheric and Local Environment team, and Defra's INSPIRE team:

David Carslaw

David Carslaw is a Principal Scientist at King's College London Environmental Research Group. His research interests are mostly related to understanding how transport systems affect air quality. Part of this interest is related to the development of analysis techniques to help better understand these linkages, particularly through the Openair project. David has been a member of AQEG since 2002.

Sue Grimmond, Kings College London

Professor Sue Grimmond joined King's College London in January 2006 after being Assistant, Associate and Full Professor at Indiana University, Bloomington USA. She completed her undergraduate degree (BSc Hons) at the University of Otago, New Zealand, and graduate degrees (MSc and PhD) at The University of British Columbia. Sue is on the editorial boards of Journal of Applied Meteorology and Climatology; Agricultural and Forest Meteorology and Advances in Meteorology. She is the 2009 recipient of the Helmut E Landsberg Award from the American Meteorological Society 'for numerous important contributions that have greatly advanced urban meteorology and urban climate sciences, and for sustained and effective leadership that has energized the urban climate research community'.

Bryan Lawrence, STFC

Bryan Lawrence is the Director of the Centre for Environmental Data Archival at STFC, where he runs the NCAS/British Atmospheric Data Centre (BADC) and the NERC Earth Observation Data Centre (NEODC). One of Bryan's areas of expertise is in information architecture for environmental data.

2.3 Acknowledgements

We would like to acknowledge the contributions of the following organisations in the production of this scoping study report:

AEA
Air Quality Consultants
Bureau Veritas
British Atmospheric Data Centre
CEH
Defra
Department of Environment Northern Ireland
Kings College London
National Physical Laboratory
Welsh Assembly Government
TRL

3 Data to be integrated

The priorities for data integration have been defined as follows:

- Top priority: NAEI; PCM; AURN; LAQN; non-automatic network data (hydrocarbons, heavy metals, black smoke); Devolved Administration archive data
- Mid priority: Local Authority action plans, review & assessment information, Local Air Pollution Control (LAPC) data, Parts A&B process data, Noise & DEM maps, deposition monitoring and modelling data
- Low priority: Regional or city scenario data as modelled through other research contracts
- Non-air quality data

This scoping study focuses on the top priority data sets, and examines how maximum efficiency and benefit can be gained through the integration of these data. Mid- and lower priority data have also been considered in the development of the proposed solution and appropriate recommendations are made to ensure that the proposed infrastructure does not preclude the integration of these important but lower priority data sets into the framework.

Although not considered by this scoping study, any future integration should also include datasets from diffusion tube networks, other non automatic monitoring networks and local atmospheric emissions inventories (including the London Atmospheric Emissions Inventory and the London Energy and Greenhouse Gas Inventory) in the mid-priority category due to their widespread use and the current difficulties highlighted by the users in accessing these data.

3.1 Top priority data

3.1.1 NAEI

The National Atmospheric Emissions Inventory (NAEI) is funded by Defra and the Devolved Administrations, and compiles estimates of emissions to the atmosphere from UK sources such as cars, heavy goods vehicles, power stations and industrial plant. The programme also provides key information such as emission factors, and estimates of industrial greenhouse gas emissions, which helps participants of the UK Emissions Trading Scheme.

3.1.2 PCM

Pollution Climate Mapping (PCM) datasets include maps of roadside concentrations and background concentrations of numerous air pollutants across the UK and exceedence statistics for Air Quality Strategy³ objectives and EU limit values/target values. These are generated by models using monitoring data, emissions inventory data, point source data and meteorological data as inputs.

3.1.3 AURN

The Automatic Urban and Rural Network (AURN) consists of about 130 air quality monitoring stations located throughout the UK. All the stations use continuous automatic monitoring equipment to record concentrations of NO_x, SO₂, CO, O₃, PM_{2.5} and PM₁₀. The data from the network are available on a hour-by-hour basis on www.airquality.co.uk and are provided by UK Government annually to the European Commission in compliance with EU Air Quality Directives. A significant amount of metadata for the AURN is publicly available at www.bv-aurnsiteinfo.co.uk.

³ The Air Quality Strategy for England, Scotland, Wales and Northern Ireland, July 2007, Defra and Devolved Administrations

3.1.4 LAQN

The London Air Quality Network (LAQN) is a group of air quality monitoring stations in London, Essex, Kent and Surrey. Each borough funds the monitoring within its own area, with the exception of eight sites in London which are funded by Defra and are affiliated into the AURN.

3.1.5 Non-automatic networks

Black smoke, hydrocarbons and heavy metals are measured at sites across the UK using non-automatic analysis methods which produce daily, weekly or fortnightly average concentrations.

3.1.6 Devolved Administration data

Concentrations of NO_x, SO₂, CO, O₃, PM_{2.5} and PM₁₀ are measured at automatic monitoring sites in Scotland, Wales and Northern Ireland, using equipment similar to that in the AURN. Although the AURN does include sites in these countries the full Devolved Administrations datasets are much larger.

3.1.7 Summary of Top Priority datasets

Table 3.1 Top priority dataset managers, location and format summary

Dataset	Managed by	Location	Format
NAEI	AEA	AEA servers including www.naei.org.uk	MS Excel and MS Access
PCM	AEA	AEA servers	CSV files and MS Excel, ESRI ArcGIS
AURN raw and validated data	BV	http://www.bv-aurnsiteinfo.co.uk (and Uploaded to www.airquality.co.uk)	From a Proprietary from the Indic AirViro software package.
AURN ratified data	AEA	AEA server www.airquality.co.uk	Proprietary software running on Open Source Technology including MySQL Database
LAQN	KCL	www.londonair.org.uk	SQL Server 2005
Non-automatic networks	NPL	www.airquality.co.uk	MS Excel spreadsheets
DA monitoring data			
Scotland	AEA	AEA server www.scottishairquality.co.uk	Proprietary software running on Open Source Technology including MySQL Database
Wales	AEA	AEA server www.welshairquality.co.uk	As above
Northern Ireland	AEA	AEA server www.airqualityni.co.uk	As Above

3.2 Mid priority data

CEH Deposition monitoring and modelling data have been considered as part of this study. These will be covered by the INSPIRE requirements and should therefore be readily integrated into any proposed air quality solution. However, it will be important and mutually beneficial for CEH to remain fully involved in the development of any future integration system to avoid unnecessary duplication of effort.

Outcomes from the **review and assessment** process are currently being developed into more standardised report formats which are uploaded to an electronic on-line register of submissions. In the future this could be further developed in order to include a directory of the underlying data files used to support the review and assessment process in each local authority. These data files would need to be in a prescribed format compatible with the Defra integrated air quality infrastructure but there is no reason why this could not be done.

Emissions data from LAPC, Parts A& B processes and so on are likely to require further work to integrate where they are not already covered in sufficient detail by emissions inventories. These data are often not easily accessible in electronic form.

Noise & DEM maps already exist and will come under the requirements of the INSPIRE Directive. They should therefore be able to be readily integrated with any air quality assessment infrastructure in the future should they be considered to be important. Much of the data that underpin the online noise maps at www.defra.gov.uk/environment/quality/noise/environment/mapping/index.htm would also be necessary inputs to future air quality tools, for example, traffic and land use spatial data.

Other non-automatic monitoring network data are managed by many of the high priority data contractors, and a horizontal approach to integration would be recommended, to prepare all datasets managed by each contractor at the same time, thus maximising efficiency and maximising the availability of data.

Local emissions inventories such as the London Atmospheric Emissions Inventory (LAEI), London Energy and Greenhouse Gas Inventory (LEGGI), and other Local Authority inventories are important datasets which would provide very valuable inputs into current and future models and tools including the PCM, and as such we recommend that they are considered for integration. The proposed solution detailed later in this report has been designed to allow for additions of datasets in the future, if the integration of these data were not possible immediately.

3.3 Low priority

No consideration of low priority data sets has been made so far as part of this study. These regional modelling studies will be investigated as to their value as part of any follow-on work.

3.4 Non air quality data

Key to the success of developing or improving the efficiency of on-line air quality tools is the availability and accessibility of essential non-air quality data sets. This includes information held by government departments, agencies and other organisations who have not been involved in this pilot study. In order to make progress these organisations needed to be included early in the next phases of the study, and the costs and barriers for integration of these data sets identified.

Some issues with key data sets have already been identified through discussions with the data providers and users:

3.4.1 Meteorological Data

Meteorological data for analysis purposes are mainly provided by the Met Office - the UK's National Weather Service. This is a Trading Fund within the Ministry of Defence, operating on a commercial basis under set targets. Requests for meteorological data from air quality specialists are therefore currently treated on a commercial basis and may incur significant expense.

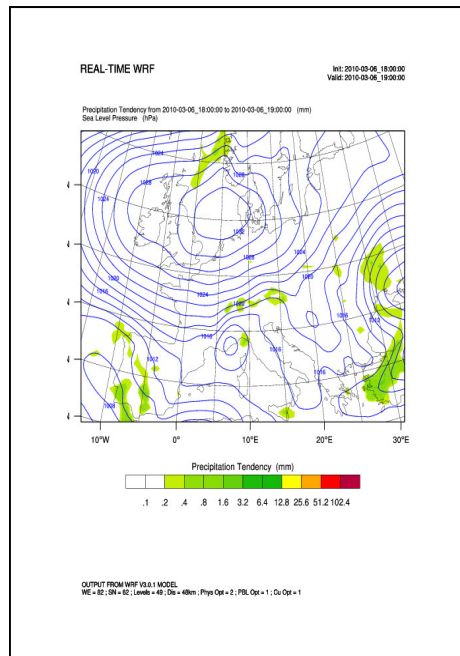
The most representative meteorological measurements for a particular study may not necessarily be from a measurement station since there is not always one in the vicinity – modelled results may therefore be offered instead.

Within these constraints the Met Office does offer the most comprehensive and quality assured data sets which are currently available for use by air quality specialists in the UK.

Assuming that there are no plans to make the UK national meteorological data sets freely available and immediately accessible to everyone, other data sets may be considered for the purposes of this study:

- Local authority air quality monitoring stations are often configured to make their own real-time measurements of wind speed, wind direction, temperature & relative humidity on 3 metre met masts. Whilst these data are obviously specific to the local situation and not rigorously quality assured by meteorological experts, it is possible that they could be made freely available to an integrated data solution.
- Data from meteorological modelling funded by Defra or others through other UK contracts – for example the UK Air Quality Forecasting or Tropospheric ozone modelling contract. Forecast data fields are currently produced on a daily basis which could be made available to other researchers. Again these data will not be fully quality assured by meteorological experts, but the Open source WRF model which is used for air quality forecasting is widely accepted for use within the air quality community and can provide output fields of the type illustrated below.

Figure 3.1 Realtime WRF model used for air quality forecasting



3.4.2 Traffic Data

This is often considered as difficult to get hold of, however our initial research indicates that information from the Highways Agency is in fact freely available as a real-time data feed or through historical data files:

The real time information required on Traffic Flows on the Highways Agency's network can be obtained on a Datex II feed. More information concerning the Travel Information Highway (TIH) can be found at <http://www.tih.org.uk/index.php/Home>. This data provides traffic volumes and vehicle classifications for all links across the network.

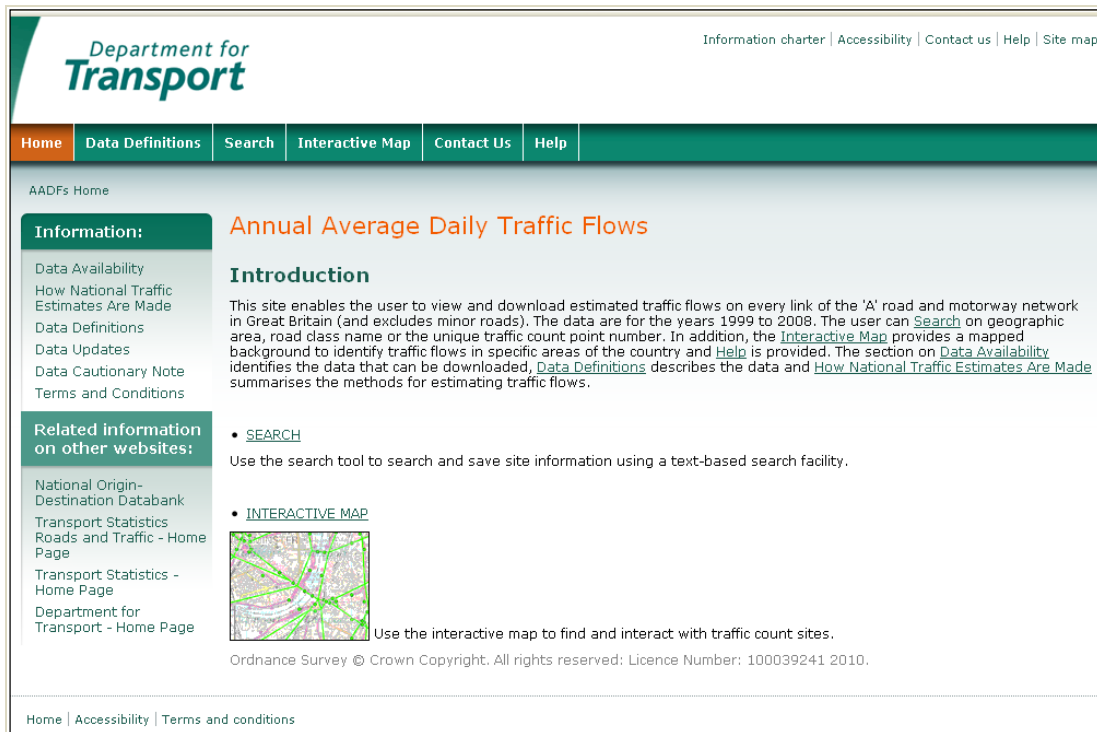
Historic Traffic Data is also freely available and access to this can be obtained at the following website <http://trads2.co.uk/>.

Annual Average Daily Traffic Flows are also available as downloadable files and through interactive maps on the Department for Transport's website at <http://www.dft.gov.uk/matrix/>.

Kings College also referenced the ANPR (Automatic Number Plate Recognition System) as a traffic data source.

The data cover all major roads i.e. motorways and A roads and excludes minor roads. The roads are broken up into a series of links with each link comprising a stretch of major road between 2 consecutive junctions with other major roads. A link may also start/end at a local authority boundary or an urban/rural area boundary. A traffic count takes place on each link of the major road network and is used to estimate the annual average daily traffic flows. These data and the associated road network information can be viewed and downloaded from the web site for each year from 1999 to 2008.

Figure 3.2 Daily Traffic Data screenshot, Department for Transport



Since these are publicly available datasets there should be no reason why through liaison with the data providers they could not be made available in a straightforward manner to an integrated air quality platform.

3.4.3 Data.gov.uk

As a result of one of the recommendations of the Power of Information Review⁴ the UK Cabinet Office has established a project to make public Government data more widely available. This initiative aims to embrace the Internet and open data standards to make information more widely available for organisations and individuals who wish to build tools with Government data. This project is currently in its infancy; it is largely a searchable catalogue of a range of different data formats (CSV, Excel, XML). Ultimately the project aims to make use of semantic web technology which will make it possible to link data together, but this will require changes to the data (for example adding additional mark-up to HTML pages to make them re-usable by systems) which are currently available so that it takes advantage of Linked Data (see Appendix 7 for explanation of Linked Data). Many of the aims of data.gov.uk dovetail nicely with SEIS and INSPIRE and it is likely (but not yet confirmed) that the data.gov.uk website will be one of the ways in which the UK complies with its INSPIRE obligations.

The data.gov.uk website and project is an important development and one that needs to be considered for any future plans for air quality data and other Government datasets.

All datasets with a spatial element, regardless of whether they are included in the data.gov.uk website, will be required to comply with INSPIRE, and therefore issues of compatibility are likely to be less of a barrier than issues of availability, which currently cause significant problems for data users, as discussed in Section 4.

3.4.4 Other datasets

Other datasets which are publicly available, owned by Defra, another Government Department or the Devolved Administrations should be considered for integration in the future. This could also include, but should not be limited to:

RESTATS, Renewable Energy STATisticS database for the UK.

The RESTATS database is currently being improved and it is planned that it will offer XML data feeds on Renewable schemes. This data could be very useful for air quality modelling – for example should certain types of renewables schemes be operating in an area (e.g. biomass) this could have an impact on air quality in that area.

ETSWAP, Emissions Trading System Workflow Automation Project

The ETSWAP project is a project planned by the Environment Agency to capture the data around EU ETS – initially for Aviation but potentially being expanded to static sources and Marine in the future.

Various Data Frameworks including English Housing Survey and upcoming National Housing Model

These data frameworks contain data on the UK housing stock and energy usage by homes separated out by region.

⁴ The Power of Information: An independent review by Ed Mayo and Tom Steinberg, June 2007

4 Current Situation

4.1 Overview

Air quality data have been captured, quality assured and checked, processed and used for modelling in various formats and for a range of purposes for over 30 years. Over that time systems have been improved and tweaked to fit changing requirements but in some cases the underlying data management technology has not evolved and moved with the times. This is partly due to the fact that the capture of a lot of these data has to be done in real time on a 24/7 basis, making major changes potentially risky and costly unless carefully planned and managed.

Three key problems are:

- The lack of structure in the architecture of the overall system of air quality and other datasets
- The disparity of the datasets in terms of standard formats
- Lack of or inconsistent metadata

Architecture

The current situation around air quality data is one typical of many systems that evolve over time without a clear pre-defined technical architecture. It is tactical in nature; meeting the specific needs of an individual project or contract rather than allowing for a more strategic view and efficient reuse of the data assets. Each of the UK's data providers, and in some cases each dataset, has a different system to capture and manage the data. This in itself is not a problem (in fact the INSPIRE Directive states that "*Data should be collected once and maintained at the level where this can be done most effectively*") but the approach to the data is not standardised and the architecture of the different systems is typically not compatible with each other.

Disparity of Datasets

As a result the datasets are currently disparate and the access to them is limited to those individuals and groups who know where to look and how they are structured. The UK has numerous separate databases and Geographic Information Systems (GIS) for air quality and other data, many of which are incompatible. It often requires manual data export and import processes or analysis processes to use the data and make them meaningful. Data integration and manipulation can therefore be very difficult, if not near impossible.

Geographical information systems are valuable tools for interpreting data with a spatial element (referring to a geographic location) such as environmental air quality data. Some of the systems used currently to manage air quality data are not full GIS systems, but they offer a broad geographical view onto the data. Others offer far more detailed resolution of their spatial properties (e.g. Pollution Climate Mapping (PCM data)).

Metadata

Metadata (data that describes data) is either very basic or is specific to the database or dataset in use on a project. This makes it difficult to quickly know if data are compatible or comparable. In addition this makes comparing measured data (for example the Automatic Urban and Rural Network (AURN), CEH deposition data), catalogued data (for example the National Atmospheric Emissions Inventory (NAEI), PCM) and projected data (NAEI, PCM projections) extremely difficult. There are therefore limits to the capability to cross-analyse or overlay data from different sources.

Examples of this include:

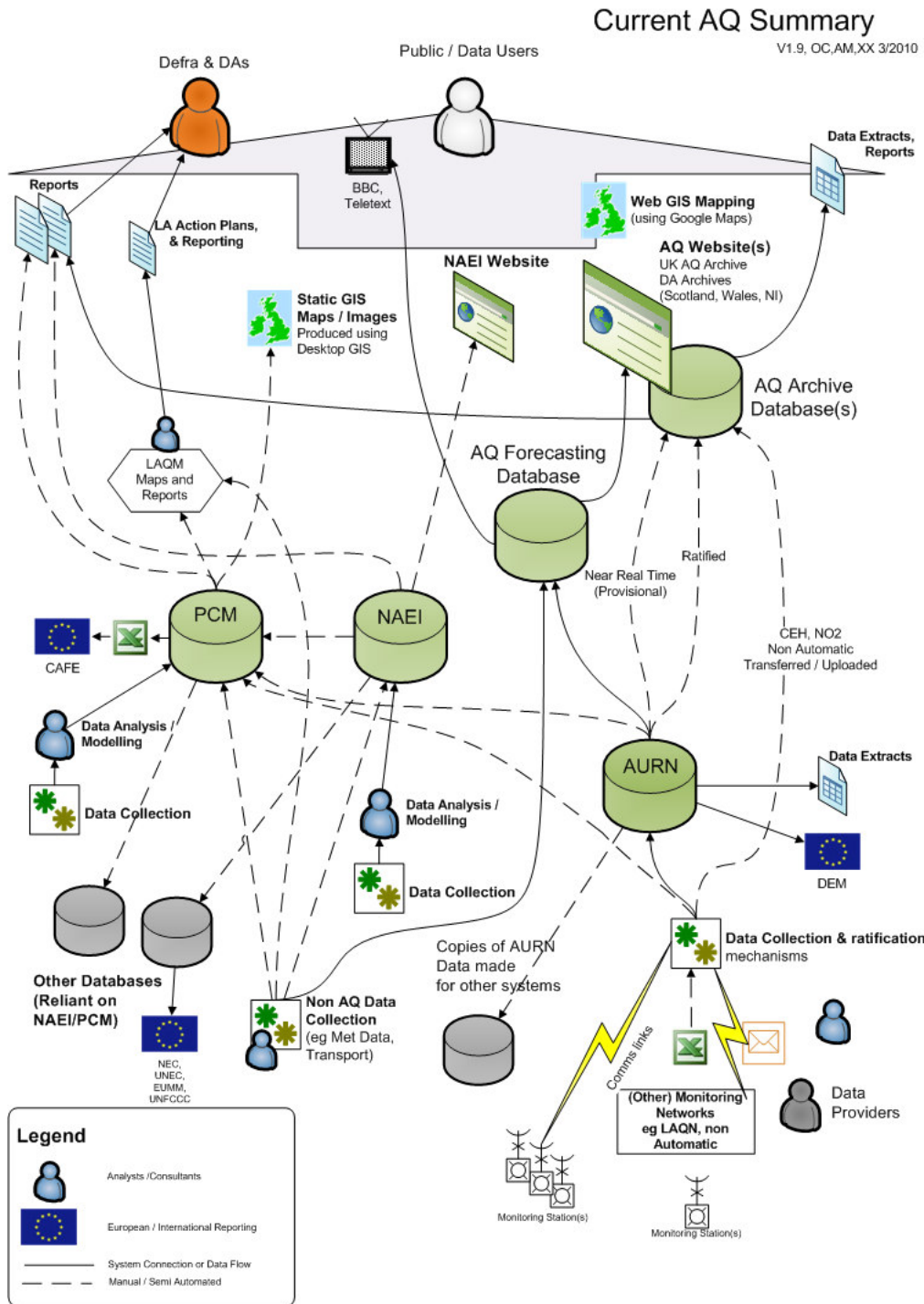
- Inconsistent naming of data attributes and identifiers.
- Inconsistent descriptors (e.g. definition of 'roadside' is different in different datasets)
- Spatial and temporal information can be stored differently across datasets

The mixture of dataset formats, metadata and platforms means that building new tools could be very costly. There is also a risk that tools developed in the current situation may not offer a robust picture of the state of UK air quality and the cost of developing such tools is difficult to estimate as a whole.

4.2 Overview Diagram

The diagram below illustrates an overview of the current situation with regards to air quality datasets. Note that it does not illustrate all datasets or data flows and focuses on the priority datasets as defined in Section 3.

Figure 4.1 National and Regional web sites



In summary it is easy to draw from this diagram that building a holistic view of air quality systems and the inter-relating variables and external factors is currently difficult to achieve.

Evaluating the current situation also highlights:

- Data duplication and out of date copies of data are being used
- There is a lack of consistency in how data is managed or exchanged
- Elements of “supplier tie in” in places as recreating some of these systems could be difficult for a new supplier

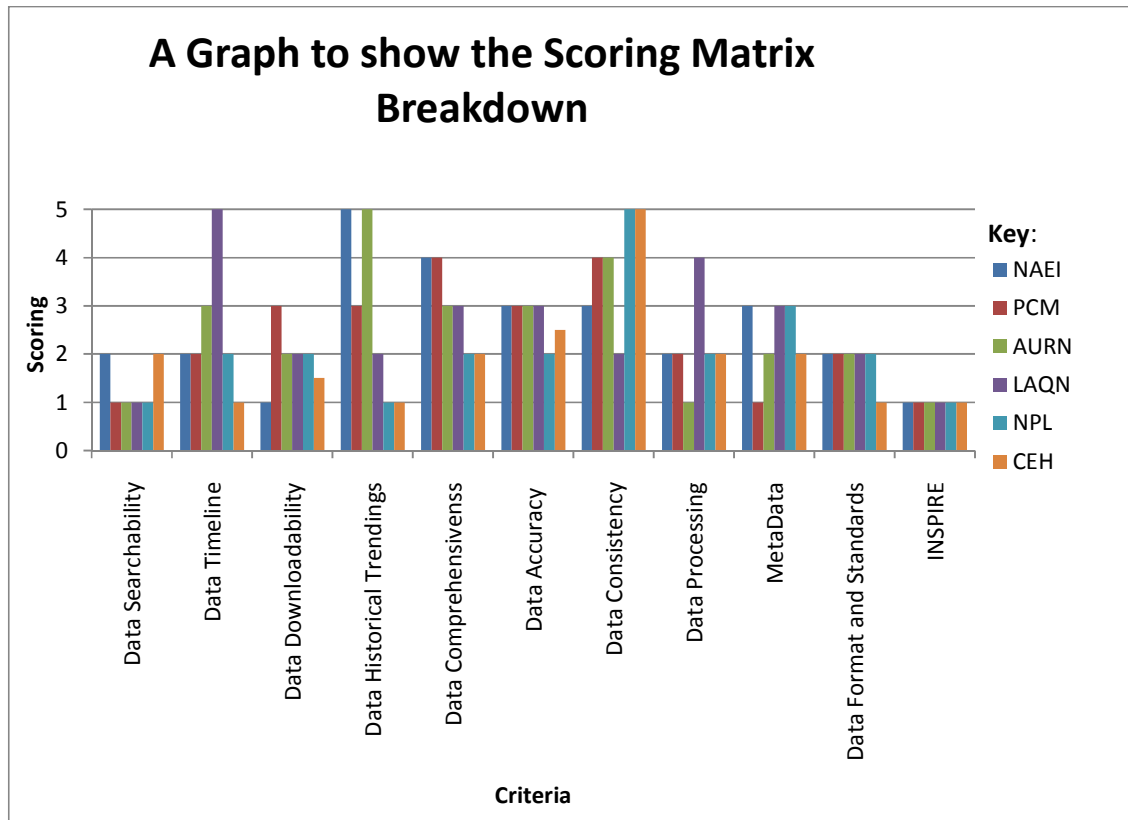
4.3 Assessment of Air Quality Datasets

Following the Introductory Workshop in January 2010, the contractors for high priority datasets were asked to provide further detailed information on these datasets through the Data Providers Questionnaire (Appendix 4) and/or one to one interviews. Once collated, this information was subject to review as part of the scoping study and each dataset was evaluated on a range of attributes using a scoring matrix. The criteria for the scoring and the summary results are given in Appendix 6. A graphical summary is shown in Figure 4.2 for the datasets.

This evaluation has assessed datasets against a new set of criteria that previously the datasets have not had to comply with. Therefore lower scores do not mean that the datasets are not fit for their current purpose, only that they will require more transformation to integrate them.

The focus is on the structure of the datasets in view of integration, rather than the potential for developing tools for analysing and reporting these data..

Figure 4.2 Scoring matrix summary results



The results from the data providers' survey identified that the main issue with the current situation is the lack of structure. There are a high number of different manual processes related with the datasets throughout the entire data cycle, from the data entry phase, to the quality acceptance/quality control phase right through to the reports produced based on the data. This means that multiple formats of data are sent to the data host, the contractors of the UK Air Quality Archive (www.airquality.co.uk), who often must manually manipulate the data to ensure that the correct data are entered into the system.

Most of the datasets are not tagged to indicate what they contain, or have a consistent standard by which the datasets are tagged, and this results in the data user being unable to find the dataset that they are searching for. It is therefore apparent that data users will have to spend time to understand what the dataset actually contains, and this conflicts with the INSPIRE principle: *It should be easy to discover which spatial data are available, to evaluate if they are fit for purpose and to know what conditions apply for its use.*

This lack of structure across the various datasets is shown not only by the data searchability scores but also from the data downloadability scores. For some of the datasets the data was only shown in report format, for instance in PDFs which causes a problem if any user wishes to use these data for further analysis. *This contradicts the INSPIRE principle: It should be possible to combine spatial data from different sources.*

As many of the datasets are stored in dissimilar formats, a user wishing to do individual analysis on the dataset would have to download it and manipulate it to conform with their standard format, for each of the datasets which they wish to incorporate into their analysis. This clashes with the INSPIRE principle: *Data should be collected once only and shared between all levels of government and all stakeholders.*

Another issue identified by the survey is that data collection methodology has changed over time, so despite some of the datasets scoring highly on the data historical trending, all of these data would have to be manipulated if any data structure principles were to be produced.

As the graph shows, most of the datasets scored highly in the criterion which just inspected the dataset individually; such as data accuracy and data consistency. However when considering how they would integrate with different datasets, especially those held by a different data provider, the scores were lower, for example for the data searchability, data processing and INSPIRE compliance.

A summary of results from the scoring matrix is shown in Table 4.1. The scoring criteria and full set of results can be found in Appendix 6.

Table 4.1 Dataset integration suitability scores

Criteria	Datasets					
	NAEI	PCM	AURN	LAQN	NPL	CEH
Data Searchability	2	1	1	1	1	2
Data Timeline	2	2	3	5	2	1
Data Downloadability	1	3	2	2	2	2
Data Historical Trendings	5	3	5	2	1	1
Data Comprehensiveness	4	4	3	3	2	2
Data Accuracy	3	3	3	3	2	3
Data Consistency	3	4	4	2	5	5
Data Processing	2	2	1	4	2	2
Metadata	3	1	2	3	3	2
Data Format and Standards	2	2	2	2	2	1
INSPIRE	1	1	1	1	1	1
Total	28	26	27	28	23	22

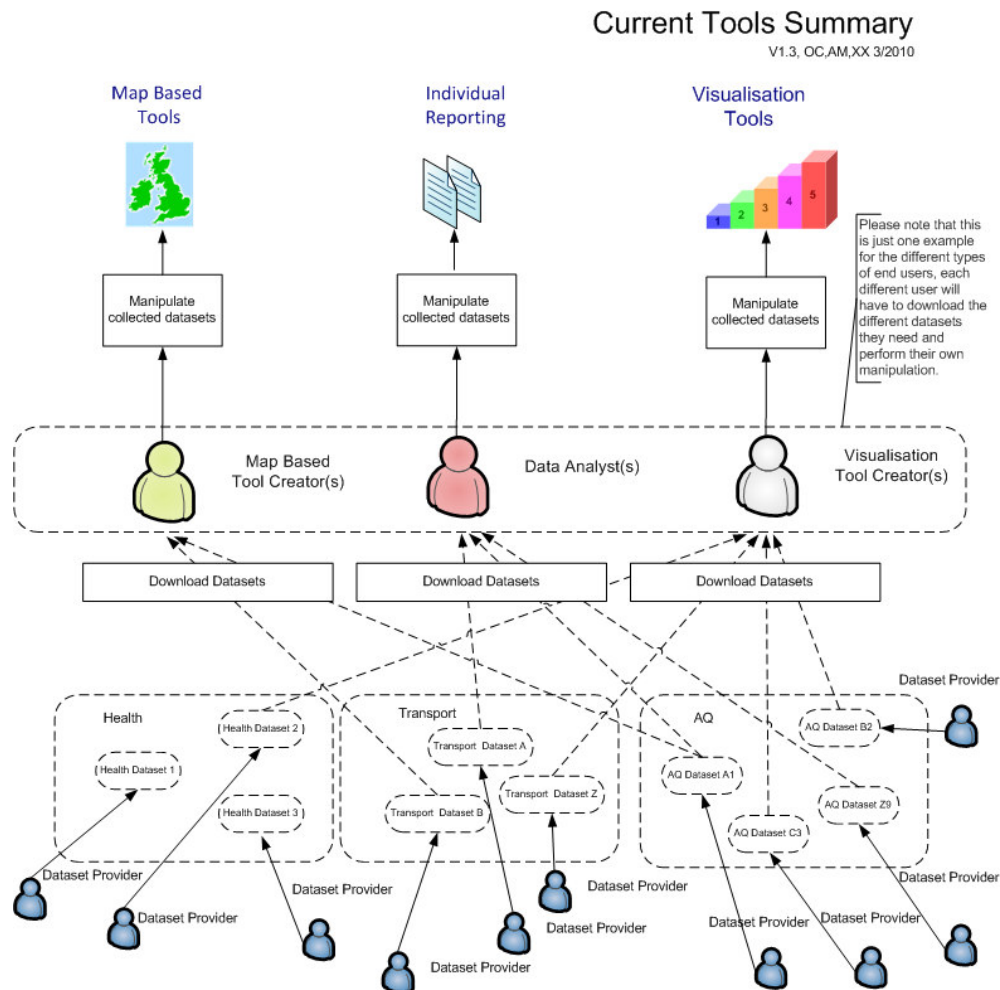
4.4 Current Architecture for Tools

The below diagram illustrates the current challenge associated with building tools that require multiple data sources. A similar picture to that illustrated in the current summary in Section 4.2, in that there is not a consistent or efficient approach to re-using data from different providers and systems. As a result many of the current tools are based on processes that involve manual data manipulation or other inefficient data management techniques.

The below diagram only shows three users of the various datasets, one for each of: Map Based Tools, Individual Reporting and Visualisation Tools. Despite only showing three users on the above diagram, it is already appearing complicated with users duplicating data by downloading and manipulating it for their own tools, for example AQ Dataset A1 is used for both Map Based Tool Creator and the Data Analyst this means that there are three versions of this dataset including the original AQ Dataset A1.

The tool creators will have to download the datasets that they wish to incorporate into their tools. So for every tool that is produced the tool creator will have to find the desired datasets, manipulate each of them into the same structured so they can be integrated together and then produce a tool to use these modified datasets to produce the desired output.

Figure 4.3 Current Tool Summary



A comparable diagram showing the future approach to building tools can be found in section 5.4.

4.5 Local Air Quality Management Tools

The tasks involved in Local Air Quality Management range from highly structured and prescriptive screening assessments through to increasingly complex and less prescriptive methods. The following table outlines these LAQM tasks and the methods and tools currently available and commonly used to carry out the tasks.

Table 4.2 LAQM Tools Summary

Task	Tools and data
<p>Updating and Screening Assessment (USA). This is low-level assessment of monitoring data and the likely local air quality impacts of a wide range of activities in a Local Authority. Assessments that show concentrations above prescribed trigger levels activate the need for a more detailed assessment.</p>	<p>Prescribed assessment methods are set out in the Technical Guidance⁵ and are increasingly shifting towards online reporting. Assessment still depends to some extent on local data defining emissions source activity levels and monitoring.</p>
<p>Detailed assessment. This includes the use of monitoring, meteorological, source activity and emissions data in conjunction with dispersion models and GIS to assess whether, where and when one or more of the air quality objectives may be exceeded.</p>	<p>Overall assessments are completed by Local Authorities or their consultants using a wide range of commercial or Open source models in conjunction with local and national datasets and prescribed assessment methods or guidance.</p> <p>Typical local datasets (held locally):</p> <ul style="list-style-type: none"> • Monitoring data at one or more sites • Transport data (AADT flows, share of HDV traffic, average speed – sometime much more detailed to include fleet profiles and stationary traffic times) usually from some form of transport model • Commercially available meteorological data from a representative site • Dispersion model (from a range available) input and output files • GIS data typically illustrating concentration isobars and the location of sensitive receptors <p>Typical national datasets (hosted on AQ Archive LAQM and NAEI websites):</p> <ul style="list-style-type: none"> • Emissions factors for emissions sources • Fleet profiles • Background pollutant concentration maps • Emission projections

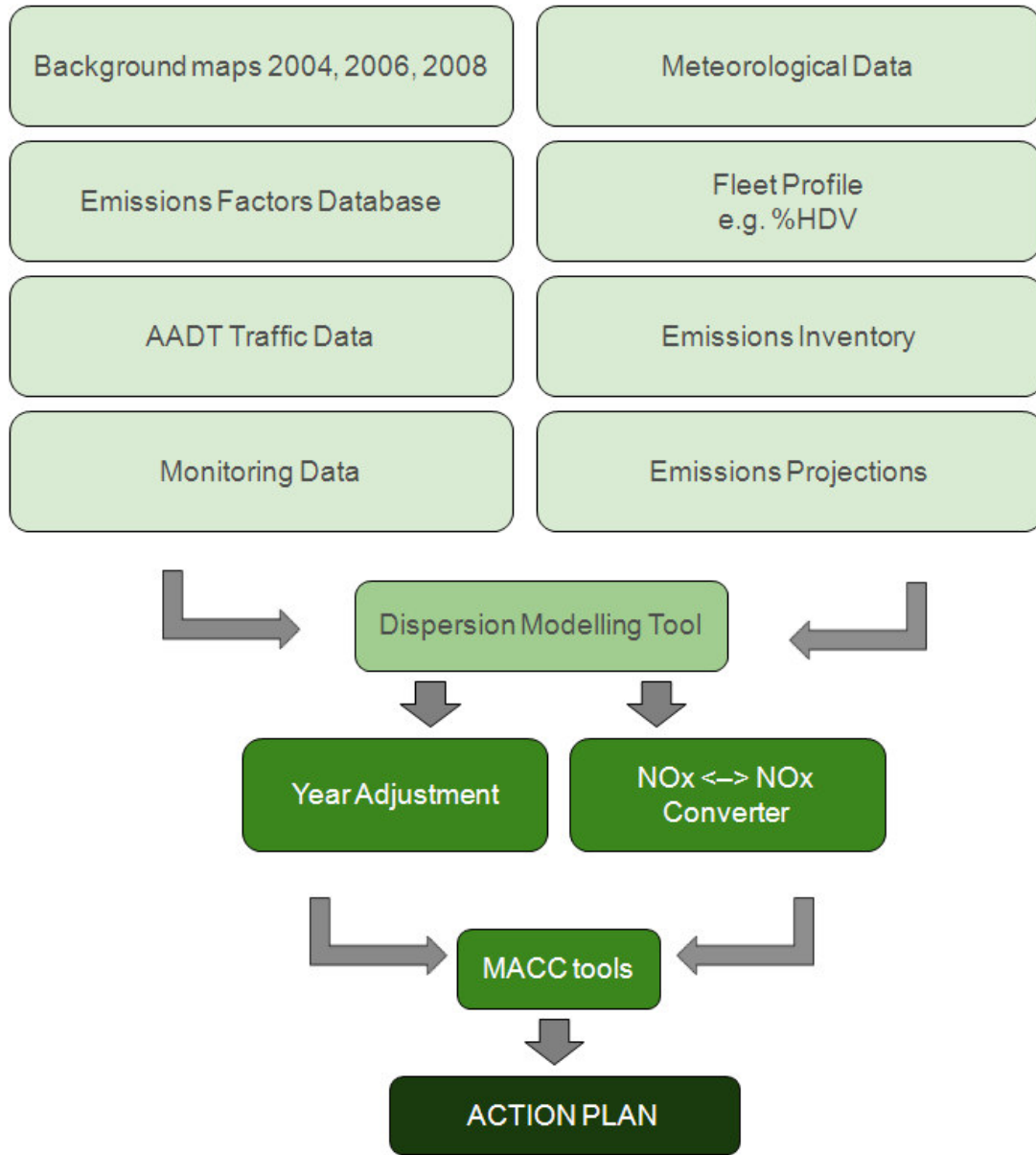
⁵ Local Air Quality Management Technical Guidance LAQM.TG(09), February 2009, Defra

Task	Tools and data
<p>Further Assessment. This is the re-assessment of those locations identified as at risk of exceeding the objectives. Typically this is a re-casting of a detailed assessment to include additional information.</p>	<p>Typically Further Assessments provide significant additional information on source apportionment and level of reductions required and many include useful scenario testing of action plan proposals.</p> <p>In some cases surveys or other data are available that disaggregate traffic flow by vehicle type or Euro standard (held locally).</p>
<p>Air Quality Action Plan (AQAP). This is the development of a strategy and action plan that is proportionate and cost-effective for the air quality issue. The plan should set out the adopted measures and when they would be implemented, project the improvements in air quality that the adopted actions may deliver and hence project when the air quality objectives may be achieved in the local Air Quality Management Area.</p>	<p>Guidance is much less prescriptive and the development of tools and data much less developed than for Review and Assessment tasks above.</p> <p>Typical local tools and data (held locally):</p> <ul style="list-style-type: none"> • Qualitative assessment of abatement actions <p>Current best local practice:</p> <ul style="list-style-type: none"> • Emissions inventories and dispersion models disaggregated sufficiently to undertake detailed cost-effectiveness and assessments (i.e. similar to those for detailed assessment but with the addition of abatement scenario analyses) <p>National data and tools (available from Defra website):</p> <ul style="list-style-type: none"> • Policy practice documents and associated data • Cost discounting tools • Damage cost accounting tools and guidance • AQAP website which collects text based examples of good practice reports, case studies and links to other helpful websites.

The most data-intensive processes are therefore the Detailed Assessment, Further Assessment and Action Planning phases, which require all the data sets (& more) identified for the national assessment, perhaps at a higher resolution and with a much more local focus.

At the centre of these processes and data flows are dispersion modelling and MACC tools as illustrated in Figure 4.4.

Figure 4.4 Example data flows for local air quality management tasks.

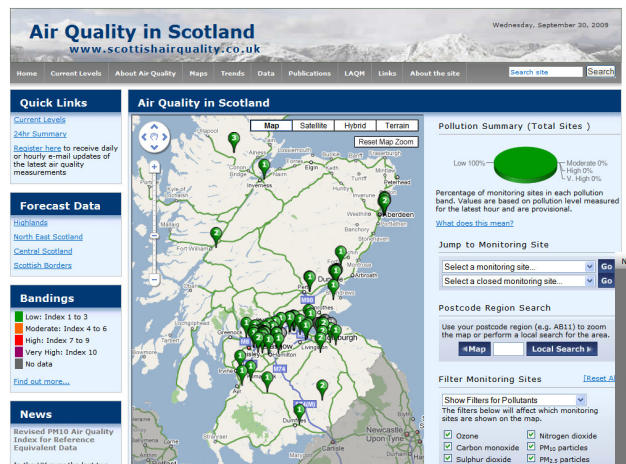


4.6 Tools for Public Information and Research

On-line air quality information is currently provided freely in a variety of formats for both public information and research purposes:

Near real-time data displays for the UK, Devolved Administrations, regions or individual local authorities use a number of tools to provide user-friendly presentations of latest air quality monitoring results & forecasts. These allow members of the public (or policy makers) to make decisions or take action based on the latest information. Actions may include taking additional medication or staying inside if you are an individual susceptible to the effects of increased pollutant concentrations, or publishing press releases and additional information regarding air quality alerts if you are a policy maker. Information may include encouraging “behavioural change” to try to reduce the impact of any current or forecast poor air quality situation. Real-time information is often provided on web-sites through colour coded mapped concentrations (using Google or other GIS tools), or through on-line graphs and displays using Open source or commercially available software. Some examples of such tools are shown in the web-site screen shots illustrated in Figure 4.5 below and Figure 4.6 overleaf.

Figure 4.5 National Web sites



www.scottishairquality.co.uk

www.airqualityni.co.uk

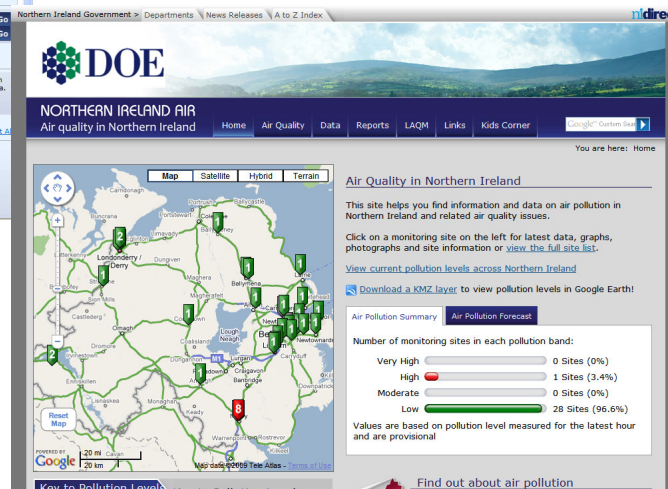


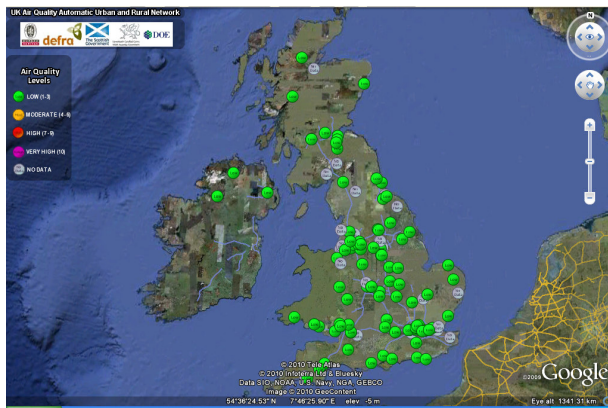
Figure 4.6 National and Regional web sites



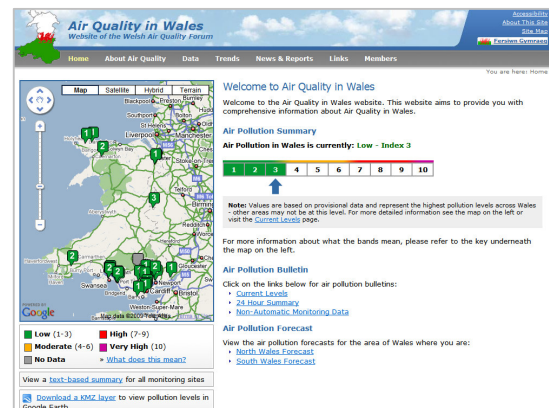
<http://www.gibraltairairquality.gi>



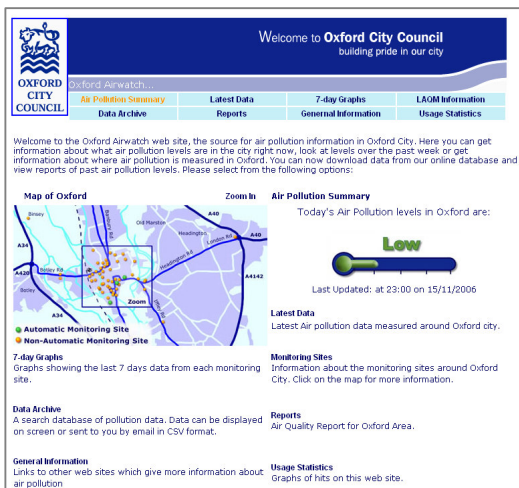
www.airquality.co.uk



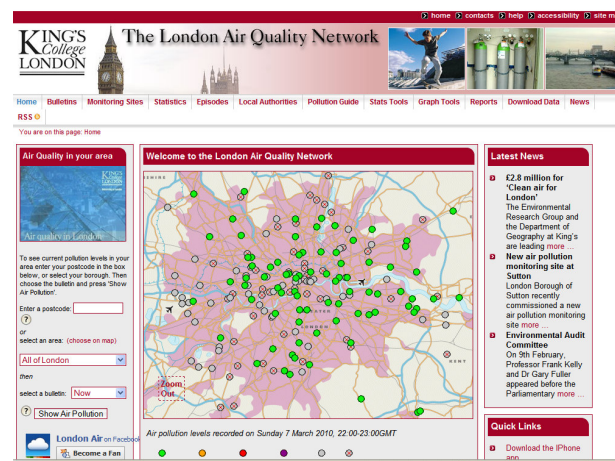
UK Air Quality data on Google Earth



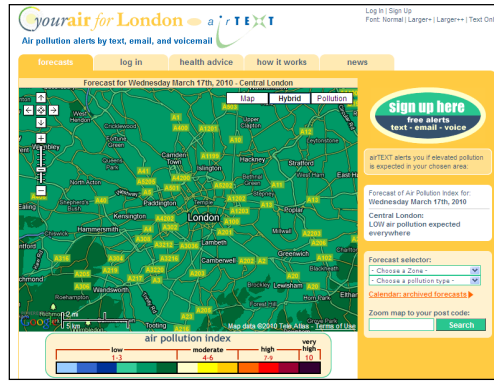
www.welshairquality.co.uk



www.oxford-airwatch.aeat.co.uk



<http://www.londonair.org.uk>



www.airtext.info/

Increasingly, tools are being developed which allow members of the public to keep up to date with latest air quality wherever they may be or whatever they are doing:

- RSS feeds, Google widgets and gadgets or iPhone Apps.
- SMS or MMS alert services
- Simplified web pages for browsing on a mobile phone or PDA.

Figure 4.7 illustrates some typical mobile web pages developed from the Air Quality in Scotland website.

Figure 4.7 Examples of Current Mobile Web Services



On-line Data Analysis and Presentation Tools allow more in-depth interrogation and investigation of data to be carried out by more sophisticated website users. Particular examples may be:

- Wind roses and pollution roses by incorporating meteorological data

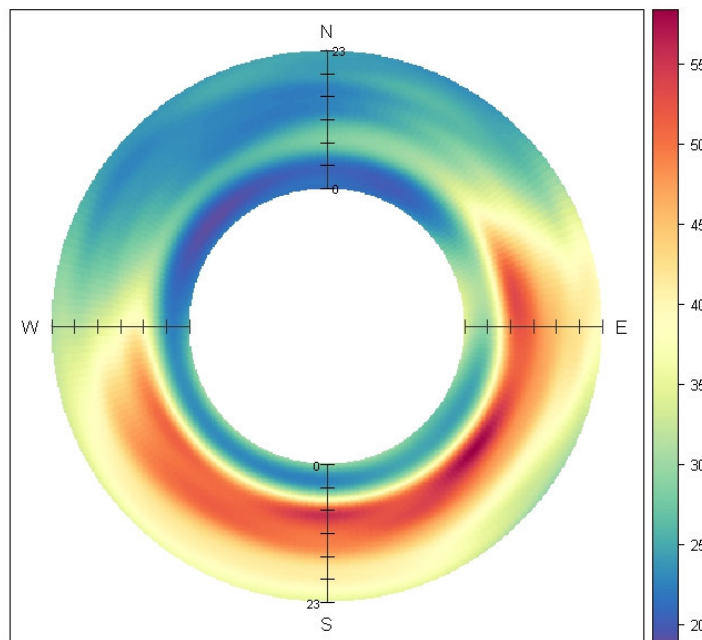
- Time series graphs
- Diurnal and day-of week analyses
- Trends analysis

One recent innovation in this area is the Open-Source Air Pollution Project which seeks to address the data user needs by:

- Developing a consistent analysis framework for air pollution data analysis
- Developing specific tools for air pollution analysis
- Making these tools freely available and Open-source
- Increasingly the 'know-how' with respect to what can be done and how to go about it

An example of the type of analysis which can be carried out using this software is provided in Figure 4.8. However, accessing some of the non-air quality data required for this type of analysis is not straightforward, and there are barriers in terms of cost and accessibility of these data sets which need to be overcome in order to release the full potential of these tools.

Figure 4.8 Example of Openair data analysis tool



This is a polar annulus plot showing diurnal variation of a pollutant concentration. This plot shows that the pollutant concentration is highest at about 9am when the wind direction is south-easterly, suggesting a pollution source to the south east of the measurement point.

4.7 National Air Quality Assessment Tools

National Air Quality Assessment tools are currently provided primarily through the Pollution Climate Mapping (PCM) contract operated by AEA on behalf of Defra and the Devolved Administrations.

These tools have been developed in order to carry out a number of critical UK air quality policy functions:

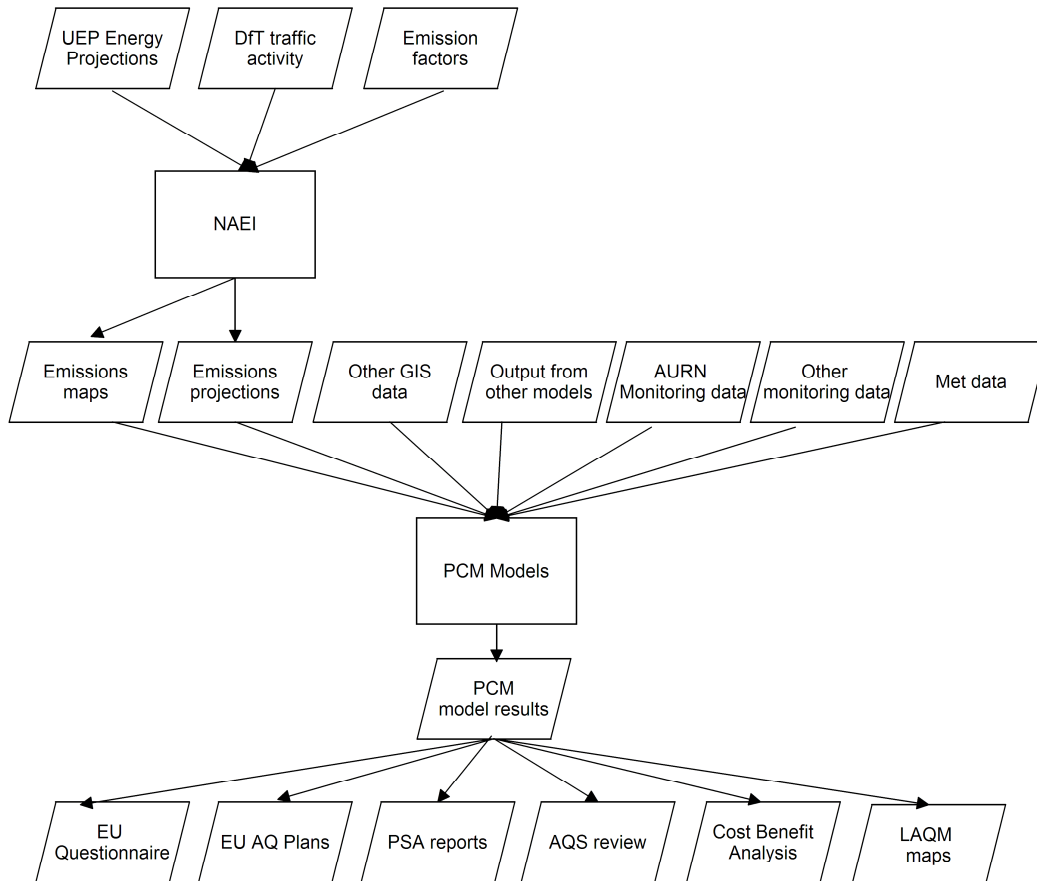
- 1 km resolution modelled air quality maps for the whole of the UK based on the most recent year of air quality monitoring results. PCM maps cover a set of predefined metrics for each pollutant across a whole calendar year

- Model near-road concentrations on the UK's major trunk roads
- Asses the compliance of latest air quality results with European Air Quality Directives
- Incorporate future emissions scenarios in order to determine the future trends and compliance situation for the UK

The tools have been developed to be operated manually off-line according to a strict annual timetable, ensuring that results are available in good time to meet the UK's statutory reporting requirements. The process requires data to be collected and input from a disparate range of sources and formats, using complex spreadsheets, databases and GIS-based modelling techniques to process and develop the model outputs. Outputs are presented through GIS compatible colour-coded mapped 1 km gridded concentrations, or through line data for the modelled roads.

PCM is an excellent example of a system which would benefit from improved integration of the input data currently required, and through enhanced GIS tools and INSPIRE compliant output data in order to maximise the efficiency and accessibility of the process and outputs. Figure 4.9 clearly illustrates the diverse range of data inputs and outputs for the PCM programme. In an integrated system the data inputs to the PCM model could be fully automated, and the resulting data, reports, plans and maps would automatically feed into LAQM and other tools.

Figure 4.9 PCM model inputs and outputs



4.8 Stakeholder Survey

AEA has sought input and feedback on the current situation from about 40 users of air quality data, and responses have been received from the organisations listed below:

AEA
Bureau Veritas
CERC
City of York council
Derry City Council
Falkirk Council
Liverpool City Council
London Borough of Hillingdon
Newry and Mourne District Council
Oxford City Council
Sheffield City Council Sussex Air Quality partnership
University of Birmingham
University of Leicester
University of the West of England

These organisations use air quality data in many ways for a variety of purposes but the most common are:

Table 4.3 Air quality data and their uses reported by stakeholder survey

<p><i>Most frequently used data</i></p>	<p>Emissions inventory maps Local and National monitoring data, specifically AURN and LAQN PCM outputs Local air quality action plans National and local fleet and transport activity data Air Quality Forecasts Diffusion tube data LAQM background maps</p>
<p><i>Common uses for the data</i></p>	<p>Model evaluation and input to models Local Air Quality Management requirements, including statutory Review and Assessment process Determining baseline air quality Pollution forecasting Statutory reporting Data for consultations Planning developments Regional strategies. Responding to public enquiries Comparison of specific sites against national objectives</p>

4.8.1 Data Availability

Only 38% of respondents who use the UK's air quality data assets have access to all of the air quality monitoring, modelling and emissions data that they require, and the majority (81%) do not have access to the associated information which is needed to give meaning to the data and to aid analysis, for instance, Met data, traffic and land use statistics, population and health data.

Air quality data

The majority of the data users source the air quality data from the Air Quality Archive with others naming other external websites including the Scottish and Northern Irish air quality websites. Some of the consultants, researchers and Local Authorities who we spoke to also commonly use their own organisation's internal databases and spreadsheets if they are available for Local or National monitoring data. Several stated that information is often requested from individual contractors because it is not publically available, but this is only possible with a network of contacts and some experience – it would not be an option for members of the public. Local Authorities and EMEP were also listed as potential sources of air quality data. A general conclusion can be drawn that the data are disparate, and some Local Authorities find data gathering so time consuming that they contract the work out to consultants. It was, however, noted that the majority of air quality data that are required are available, but only if you know where to look and often only by making a specific request to the data management unit.

The individuals who responded to our survey stated that they might themselves spend 10%-25% of their time gathering, formatting and analysing data. For some organisations the amount of time spent is often greater than one full time equivalent person each year.

Some datasets were sought after but were flagged as difficult or impossible to find, in particular, but not exclusively, exceedance data, diffusion tube concentrations, Local Authority data, hydrocarbons concentrations, carbon monoxide levels, national integrated road transport carbon and air quality emissions.

Non-air quality data

The survey also identified a lack of availability of non-air quality data that are necessary to provide context and relationships between the causes and impacts of air pollution. Many individuals responding to the survey identified the same shortfalls and listed the same type of datasets that are essential to get the most value out of the UK's air quality data assets. Specifically, these were:

- Meteorological data and cloud cover: The cost of acquiring Met data was flagged up by most respondents, and it was suggested that the air quality community could start to build a central repository for Met data collected by Local Authorities, Defra and the DAs from the Met instruments in their own monitoring stations. Although not as accurate as the data collected by the Met Office, this would be a vast improvement on the current situation. The main problem with using local Met data is ensuring that it is representative of the area where the air quality monitoring is taking place, for example, you can accurately measure wind direction in a street canyon, but it is only applicable within a very small area. ***The availability of Met data is seen by data users as a very high priority.***
- Land Characteristics including topographical data. There are restrictions on the use of the best UK land use datasets such as those owned by Ordnance Survey.
- Vehicle Statistics, in particular local bus fleet, traffic and congestion data. It was noted that these data are sometime available but in many cases are out of date or need a significant amount of work to manipulate them into a useable and meaningful format. ***The availability and integration of traffic data is seen by users as a high priority.***
- Population and health statistics.

On the whole, the individuals and organisations who were targeted during this exercise are experts in the field of air quality, often with many years experience locating and analysing air quality data. As such we can assume that the results of this survey are not representative of the difficulties experienced by a large population of people who use the data less frequently and do not have the necessary contacts in the field.

4.8.2 Data Analysis

Comments on the availability of data and analysis tools ranged from inadequate to excellent; however, by drilling down with specific questions, there is a clearer consensus of opinion.

- **81%** would welcome a single air quality portal with improvements to the availability, format and integrity of these data
- **75%** would welcome basic or complex online tools to help analyse these data
- **63%** would welcome basic or complex online tools to help present these data in visual formats. It seems that some Councils and organisations already have good visualisation tools and would rather use their own, but for the majority of people presentation tools would be very useful, especially automated data plots which would allow the user to quickly review before exporting the raw data into their own systems for further bespoke analysis.
- **56%** would welcome basic or complex online tools to help make decisions regarding air quality in the UK. This question was not applicable to all respondents but we would like to draw attention to an important point voiced by one stakeholder:

“In making decisions on air quality we would caution against the setting of an automated response, which undermines the technical excellence of air quality expertise in the UK”.

The application of local knowledge and expert judgement is invaluable and cannot be replaced.

- **69%** would welcome basic or complex online tools to help action planning. This question was not applicable to all respondents and only one organisation stated that they would not use such tools. On the whole action-planning tools would be supported.

“The lack of coherent impact assessment methods have undermined the action planning process and weakened the LAQM regime in respect of actions to improve air quality”

- However, only **44%** would welcome more centralised, automated and uniform tools and methods for data analysis and reporting. Although this would give better consistency for comparison across regions, it was noted that an automated process would make it difficult for the user to make independent checks of the data inputs and outputs and that incorrect conclusions may be drawn if the analysis method wasn't adequately understood. One key consideration should be the potential for very prescriptive tools to stifle innovation and learning, leaving the UK worse off in the long run.

Reassuringly, only a minority of data users have experienced problems with current spreadsheet-based data and tools – the main issue is that the data formats, availability and tools are not sufficiently developed, rather than the ones we have being incorrect or of poor quality.

4.8.3 Specific issues

Data availability

The primary difficulty is knowing what data are available in the first place and then being able to find them. Better search and indexing tools are essential to allow users to see and search the datasets.

Data quality

The reliability of the data and the completeness and accuracy of the metadata is of extreme importance to some (but not all) data users. It is often unclear whether or not the data quality is of a high enough standard. As such, ratification tools would be very valuable, and where data quality cannot be guaranteed, this should be flagged.

Data formats

Lack of consistency in the format of the data means that often significant re-formatting and manipulation are required to prepare the datasets for uses other than the primary purpose – usually statutory reporting. The most stated issue with the current air quality data in the UK is the amount of time taken to gather, format and harmonise the data.

Tools are inadequate

Opinion on the need for new tools and the potential take-up of them is mixed. In general the available tools are thought to be slow, although more than one respondent commented that Openair has significantly improved the situation already.

Version control and governance

These are key issues – particularly the withdrawal of out-dated tools and raising awareness of the potential issues. There is a risk that users download tools but do not check back for updates, and a suggestion has been made that it would be prudent for users to be required to register for updates upon their first download.

4.8.4 Data Analysis Tools

One of the benefits of a fully integrated system for air quality data would be the easier application of online tools to view, sort and analyse the data to give them meaning and make them applicable to real life. Initial discussions with Defra, the Devolved Administrations and the stakeholders present at the first data workshop, indicated that the tools that would be of most use fall naturally into six categories. These categories were presented to our group of data users and their suggestions are discussed below:

- **Local Air Quality Management or national air quality compliance tools.** Compliance tools would seek to simplify and speed up the mandatory reporting of air quality for Review and Assessment and to the European Commission, complimenting existing tools and further automating the reporting process. At the national level, reporting is currently done via the Data Exchange Module (DEM) and Questionnaire, both of which are very time consuming and use inappropriate technology. The development of better reporting tools would reduce errors and save both time and money, and could potentially be rolled out across Europe.
- **Presentational tools.** For the majority of people, in particular members of the public, simple presentation tools would be very useful, especially automated data plots which would allow the user to quickly review before exporting the raw data into their own systems for further bespoke analysis. Defra and the Devolved Administrations are currently supporting the development of an Openair ⁶toolset, a NERC-funded knowledge exchange project for the air pollution community. Some of the more basic Openair tools will shortly be available for public use on the Air Quality Archive.

⁶ <http://www.openair-project.org/index.php>

➤ **Practical tools.**

- Ratification tools would be useful to allow Local Authorities to ratify data more in alignment with the AURN, thus improving data quality and consistency. The ratification process is complex, requiring a number of inputs from routine calibrations, instrument services, qualitative information and expert judgement. Ratification tools will therefore never be able to substitute manual data manipulation; however, simple tools could be developed to automate parts of the process.
- Simple conversion tools would be welcomed to quickly convert concentrations into European information metrics (e.g. AOT 40).
- Defra and the Devolved Administrations, Local Authorities and Network Management Units would also benefit from asset management tools to track the deployment, age and condition of equipment, thus aiding planning for deployment and capital expenditure.
- Cross referenced monitoring and forecasting data, so users can compare what the forecast was on a particular day with the actual monitoring concentrations (perhaps using a calendar interface)
- A data catalogue or data dictionary ought to be one of the first tools to be developed after or simultaneously with the integration process.

➤ **Action planning and impact analysis tools.** Action planning tools, in particular quantification and cost benefit analysis tools would be welcomed by Local Authorities. Such tools should deal with commonly implemented measures to avoid duplication of effort each year by the same Local Authority as well as between Authorities. A tool which easily creates econometric datasets for a variety of purposes – climate change as well as AQ policy development – from existing emissions, air quality and source characterisation datasets has been suggested by one data user. Additional tools to allow realistic cost-effectiveness and other econometric analyses at the national or local scales would also be possible, leading to the integration of these data and tools into maps to facilitate cross Government policy development and to support strategic development and transport planning at the local level.

➤ **Emissions scenario testing tools.** Such tools would consist of a set of general assumptions and a set of inputs which can be selected by the user to generate a future scenario of air quality in a certain area based on these inputs. Emissions scenario tools would be used to aid policy making at the local and national level and to improve understanding of the cause and effect relationships involved.

➤ **Analysis tools.** It is envisaged that analysis tools could be developed to investigate the relationships between air quality data and non-air quality data such as traffic or health statistics, to link pollution sources with measured air quality and the impact on the population and ecosystems. However, at present the availability and integrity of non-air quality data is insufficient to support such tool development so we recommend that such tool development is reconsidered at a later date, when the data integration process is nearing completion and the necessary high and medium priority and non-air quality datasets have been integrated.

We commend greater use of Openair and other similar OpenSource software packages to develop tools. This will have the benefits of continual peer review and minimisation of costs of tool development for Defra and the Devolved Administrations.

One of the overwhelming benefits highlighted by this user survey is that the development and implementation of tools for Local Air Quality management would provide a consistent approach across the whole of the UK Local Authorities. However, the data integration process and subsequent development of tools should not replace the raw 15-minute or hourly averages that are currently available for researchers.

All tools would need suitable training documentation to avoid misinterpretation and incorrect conclusions being drawn from automated analyses. There is also a worry that the overuse of prescriptive tools removes the logical thought process, detaches the user from the data processing and cannot take into account specific attributes of each unique situation.

5 Proposed Future Approach & Architecture

Before describing our proposed future approach it is important to consider the context outlined in the previous sections of this document, which clearly show that the integration of air quality data in the UK would be complex, with a large amount of existing data and existing systems to consider and manage. Solving these problems requires coordination and standards between the data providers and data users and an underlying architecture that supports this approach. We outline a roadmap and a transition scenario in Section 6 that illustrates some shorter term quick wins. In this section of the document we are more forward looking – the possible picture in 5-10 years' time.

The approach and architecture needs to meet the requirements of integrating and standardising air quality data sets in order to overcome the challenges faced through the current situation. The future approach should also comply with future legislation such as INSPIRE and initiatives such as SEIS and data.gov.uk. The technical requirements for these are summarised and documented in Appendix 7.

5.1 Changes Required

The following changes need to be implemented to air quality and other data systems:

- Use of standards for underlying data storage formats (e.g. XML). An interim, cheaper and quicker alternative could be to specify a common data transfer method (e.g. the use of a CSV format that is already in wide use)
- Adoption of standardised metadata
- Standardised approach to linking data to a point in space (spatial) and time (temporal)
- Use of standardised approach to linking data to other datasets
- Use of reusable services (Service Orientated Architecture - SOA) that can be combined to create more complex composite services and flexible end user tools

It is very likely that for data providers it will require

- Updates to the underlying technology used to capture, process, store and disseminate the data
- Data that are being collected to be stored differently – so it complies with standards
- Data that have already been captured to be transformed to new standard compliant formats (or development of mechanisms to transform these data on demand)

It should be highlighted that it does not and should not require:

- The adoption of a single specific IT vendor's technology (for example a specific Database product)
- The creation of a single data *storage* point for all air quality datasets. Although a single air quality portal for access to air quality tools and data (that draws on many datasets) has been identified as highly desirable. However an alternative option for implementation could be to aggregate this information at a higher level, for example through an EU air quality portal or a UK environmental data portal.

5.1.1 Costs and time required for dataset providers

Costs for implementing these initial changes for dataset providers will vary depending on the providers existing infrastructure and capabilities but are likely to be in the range of **£25K to £75K** per system requiring between 3 and 12 months for implementation. Note that these costs assume that the formats and metadata have already been defined and agreed and all that is required is to implement these changes. It should also be noted that once INSPIRE and other standards mature changes may need to be made and these may of course have cost impacts. Additional costs will be incurred in the

development of tools on top of the new platform but these should have a very compelling return on the investment, particularly those targeted at Local Authority Action Planning.

5.2 Underlying Principles

The combination of the approaches outlined in Appendix 7 allow for datasets to be linked to other datasets and plotted and linked via the spatial and temporal attributes.

The future approach to integration does not necessarily mean consolidation into a single data warehouse. Whilst it may be desirable to hold a large amount of air quality data in one place to allow for efficient analysis the SEIS principles require collection locally but access globally. This approach allows each dataset provider to host their own dataset locally as a node. However this node can be accessed by all users and the dataset provider can access any other node, thus allowing the dataset to be collected locally but accessed globally. Therefore the overall architecture should be designed to support this goal. This also allows for the present situation of multiple contractors collecting air quality data to continue but it is supported more effectively through the use of a common approach and data standardisation. It also supports the ability to build a range of tools that reuse the data for new purposes.

Through the use of industry standard formats and definitions of metadata the proposed architecture will allow for the system to be modified to meet changing requirements rather than fixed systems designed around specific requirements from a set era.

5.2.1 Service Orientated Architecture

The proposed implementation route for sharing data across contractors and systems is to make extensive use of SOA. This approach abstracts the specific technology used and provides re-usable services that can be called on to prevent duplication of data across systems. SOA will only work if other factors on standardisation have been implemented – so that services are compatible with each other.

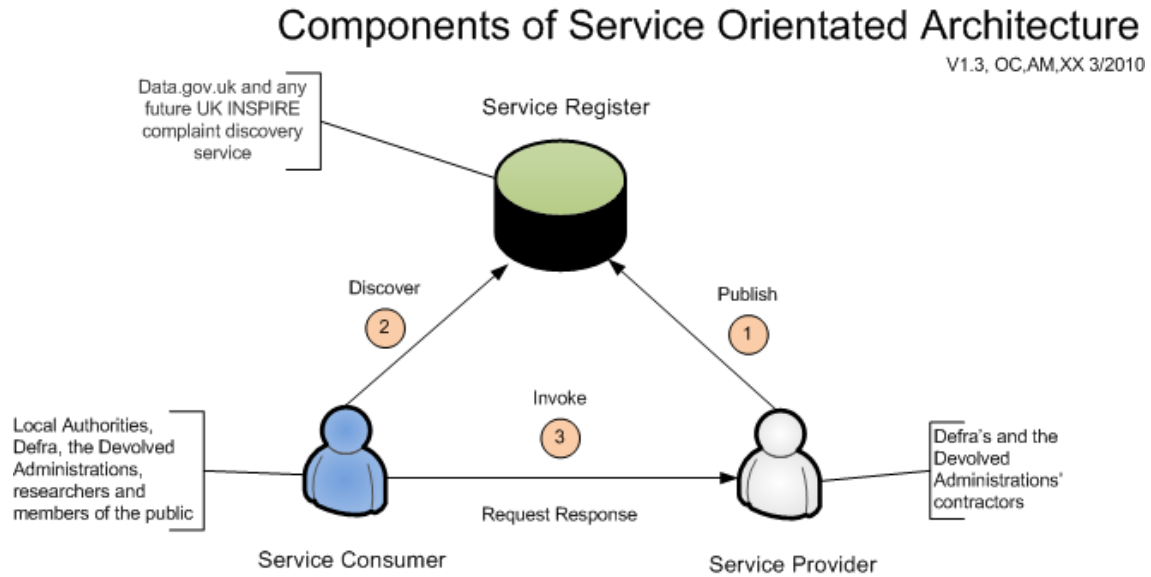
The communication can involve either simple data passing or it could involve two or more services coordinating some activity. Some means of connecting services to each other is needed.

SOA's consist of the following three components:

- Service provider
- Service consumer
- Service registry

A simple example of this in context: the service provider is Defra's and the Devolved Administrations' contractors. The consumer is any data user – Local Authorities, Defra, the Devolved Administrations, researchers and members of the public. The service registry is the proposed future catalogue of datasets that allows users (and ultimately systems and future tools) to look up what data are available and find them. One option for the service registry would be to take advantage of data.gov.uk and any future UK INSPIRE complaint discovery service that is made available. Should a centralised discovery service not be made available then provisions for an air quality INSPIRE compliant Data Discovery service will have to be considered.

Figure 5.1 Components of Service Orientated Architecture



In the diagram above Service Consumer and Service Provider refer to the IT systems being used by these respect parties.

The systems that will most benefit from the new approach are the AURN, PCM and the Air Quality Archive – as data collection and management will be based around a set of standards these systems can be greatly refined.

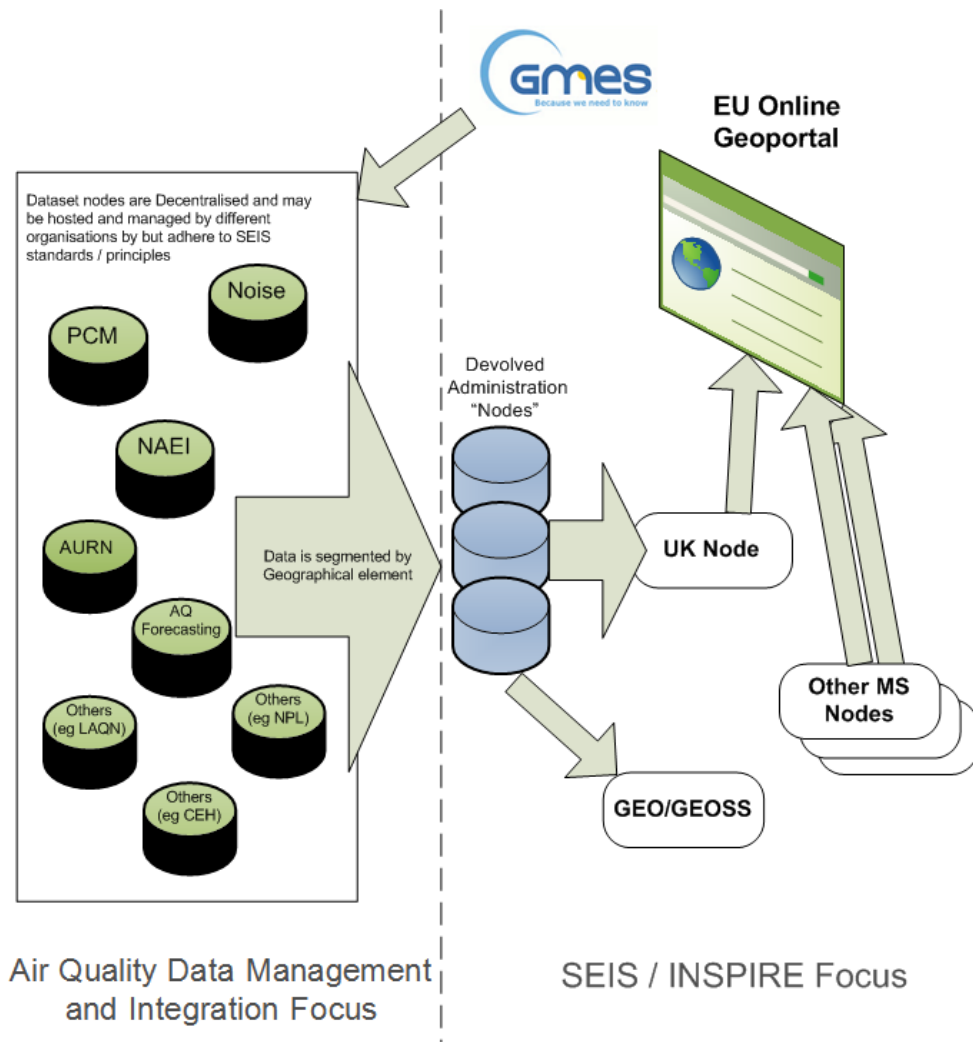
Whilst there are also benefits for the NAEI (in terms of capturing some data and also how the results of the project are report to the EU) this project has to support the capture of data from a wide range of disparate sources in a range of industry sectors including the private sector (who are not legally mandated to use INSPIRE). This makes it difficult to control the means and formats these stakeholders submit data without some additional legislation that applies to these data providers.

5.2.2 Decentralised SEIS approach

As previously discussed the SEIS approach allows datasets to be collected locally but accessed (or aggregated) up to a global level. Therefore the current mix of suppliers does not need to change – the suppliers will however need to move to a “node” based approach. Where an individual suppliers system acts as a node providing relevant data to other nodes. Further nodes can then be constructed at a Devolved Administration level (e.g. a DA air quality Portal) and then into the UK air quality node / portal. Once SEIS and INSPIRE are implemented across the EU this air quality node is then used alongside other EU member state nodes to build the complete picture for the EU. This complete picture can be accessed via the EU Online Geoportal.

In addition it allows for easier comparison of data across different Member States.

Figure 5.2 How SEIS allows for decentralised nodes



5.3 Overview of proposed future Architecture

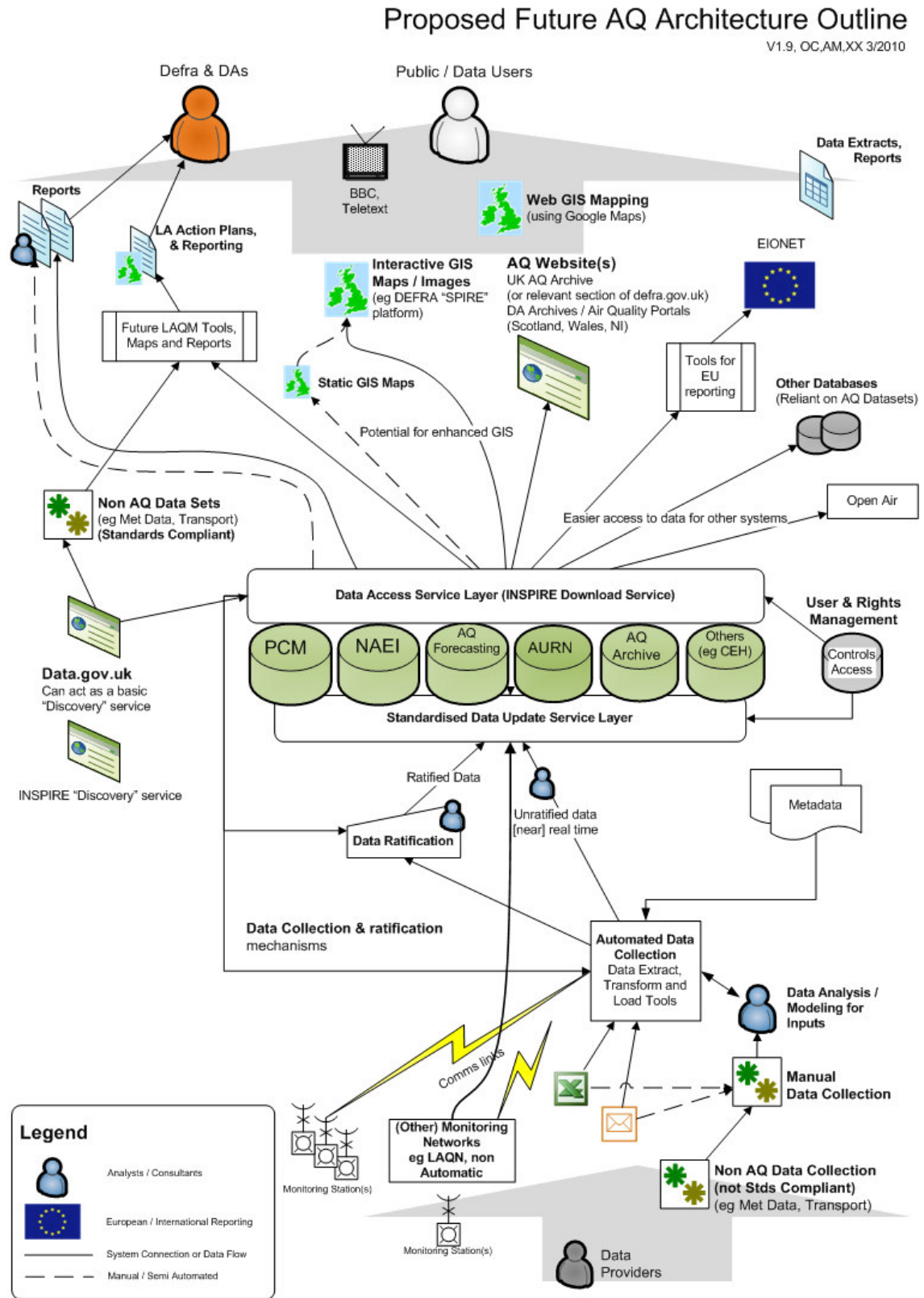
Based on our knowledge of the requirements for an integrated system and the other factors that need to be considered we envisage a revised architecture that encompasses:

- SOA – that abstract the full details of the underlying database by creating services that handle dataset access and updates / changes to datasets
 - Data access service layer (implementing INSPIRE download services)
 - Standardisation of data updates (potentially through a service layer)
 - Automated data collection is made more feasible and can be deployed where practical to replace manual data import
 - Use of ETL (Extract, Transform and Load) tools and techniques to ease and standardise and automate import processes
 - Preference for data sources to directly update the datasets via the data update service. (Over time as data sources are improved this alternative more direct and efficient approach can be taken).
- Registries that allow for human and system discovery of available datasets and services
 - Via data.gov.uk
 - Via an INSPIRE compliant discovery service
- Controlled access to data and dataset update services through user and rights management

Based on the information captured during the scoping study we have outlined how the architecture may look in the future for an integrated system. The main advantages of this approach are:

- More automation and less manual intervention. Less time will be spent searching for data and manipulation of data will be quicker due to standardised formats and new tools
- Reduced operating costs
- Robust platform for developing new products and services to add value to existing data
- Simpler and more automated reporting lines
- Ability to link or overlay data from different datasets
- Better, informed decision-making, using a wide range of datasets

Figure 5.3 Proposed Future AQ Architecture Outline



5.4 Platform for future tool development

Compared with the current situation for building tools as illustrated in Section 4.4 the future situation would provide a far clearer more efficient way to build tools that use data from a range of sources. Tools can take advantage of re-usable services and compatible datasets to access and present data in an easier to disseminate manner.

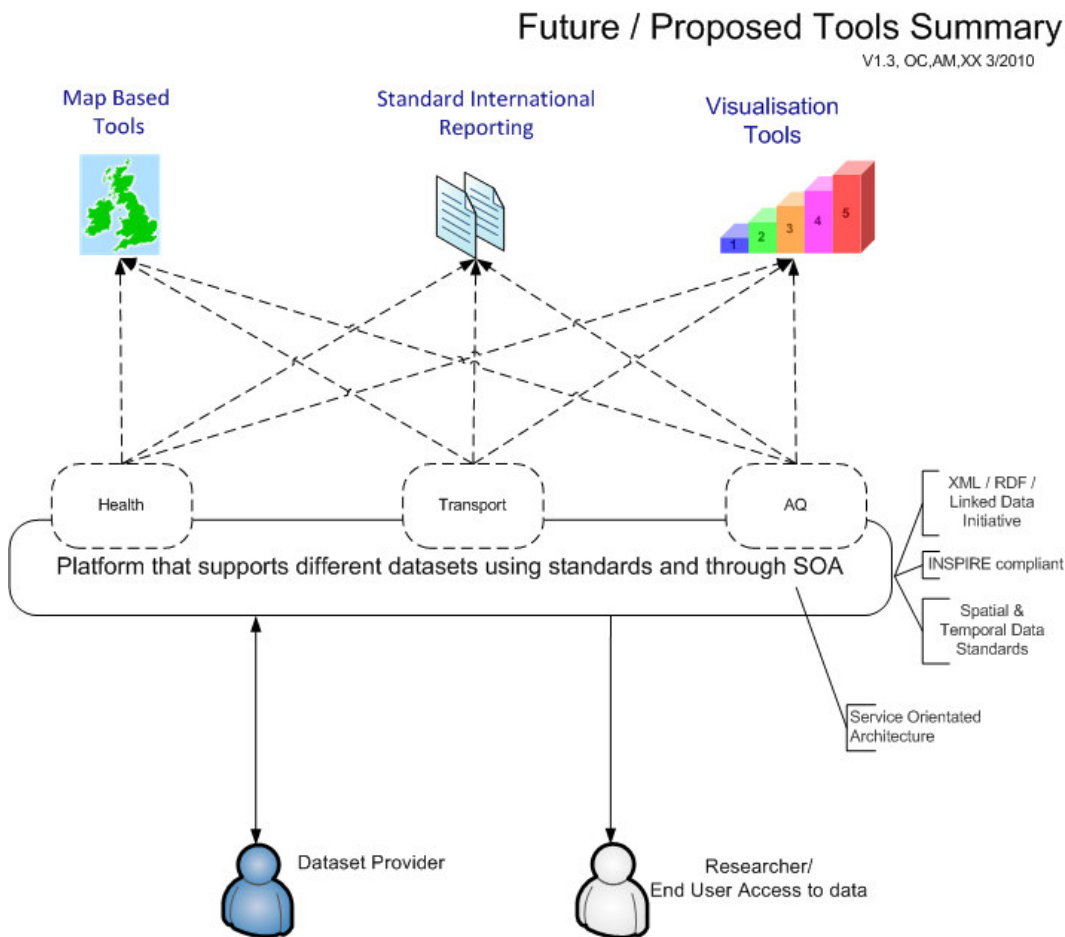
Investment in advanced and robust tools can be made based on this platform safe in the knowledge that a future proof, adaptable and standards compliant approach is being used.

The situation is slightly more complex when datasets that have not been standardised are introduced but over time, as systems and datasets are made compatible, the platform will become more and more powerful and effective.

Figure 5.4 is an example of the proposed structure, showing how data from different datasets could be used together in different types of tools. In this example non air quality datasets for health and transport data are included, but this structure would equally apply to other non-air quality data on climate change, Met data, land cover, ecosystems, population etc.

This platform is ideal for future tool development but could also be applied to tools that are currently in use, for example the Openair toolset, Volatile Correction Model⁷ and other LAQM tools. Although the proposed robust data platform is an excellent starting point, future tool requirements may highlight other data that need to be captured in the integration system.

Figure 5.4 Proposed Tools Summary diagram



⁷ www.volatile-correction-model.info/

It is important to note that there are different timeframes of data availability. For example, the AURN data is updated on an hourly basis but this is not ratified until 3-6 months later. Non-automatic data have daily, weekly, fortnightly or monthly resolution, but are not available in real-time and are currently (and likely to remain) available for public use in batches, often a few months after the sample was collected. NAEI and PCM data are available on an annual basis with about an 18 month time lag. There are, therefore, limitations to the future tool development. It will not be possible, for example, to look at high concentrations today and go straight to emissions data to see the source of the pollution.

5.5 Challenges

This proposed data integration process is not without challenges and we envisage that the most prominent will be:

- Agreeing interchange standards with other Government Departments and Agencies
- Making centralised infrastructure available to centrally manage stakeholder access and rights management
- Making sure the data are current and up to date
- Re-coding separated datasets, that is, datasets which historically were the same but have been modified over a period of time to the needs of the user
- Keeping the catalogue of datasets up to date
- Description of historic data, which may have been lost over time i.e. periodic takeover/ legacy lost (where data from legacy systems is lost due to lack of documentation/the owner of legacy system leaving)
- Different geographically based data, i.e. comparing datasets which is measured in 1 x 1 km grids with those that are measured in 50 x 50km grids. Also those measured using different co-ordinate systems (the INSPIRE regulations state that all data should be formatted to 4 different types of co-ordinate systems thus it will be easier to compare as they will have common co-ordinate systems and conversions will be possible where required through co-ordinate transformation services)
- Moving older proprietary databases towards compliance with newer IT standards (in many cases it will be more effective to replace these with new systems but this will have a more significant impact on cost).
- Overcoming IP and proprietary data issues (such as availability / cost of Met data)

5.6 Implications for stakeholders

Going through the changes that are required will affect the users in the following ways:

5.6.1 Dataset Providers

- Time – the organisation responsible for the dataset will have to spend time to alter their datasets so that they conform to the new datasets regulations
- Money – there will be an initial cost to the provider of the dataset as they will have to pay for the dataset to be altered
- Modifications to tools – some datasets providers will have proprietary databases whose sole purpose is for reporting. If the data have changed format then these reports may be generated but not function correctly or produce the correct response. These tools will have to be amended to facilitate the changes to the datasets
- INSPIRE compliant – their datasets will be INSPIRE compliant after they have modified their datasets to these standards. The dataset would have to legally confirm to the INSPIRE regulations in the future, so time and money would have to be spent on this anyway, however by conforming to these standards there is a support to help them confirm to the INSPIRE regulations

5.6.2 Dataset Users

- Reduced search time – as there will be a data dictionary containing information on available datasets
- Reduced modification time – as the datasets will be in a consistent form, less time will be spend on modifying/converting the data that is downloaded
- Reliability – the user will be able to check the reliability of the dataset that they are about to use, and check that the source is trustworthy. This is one of the problems that were reported from the questionnaires
- Compatibility – the user will be able to combine different datasets as they will be confirmed by the same standard
- Up to date data – the data dictionary will contain information regarding the last time that the dataset was updated/ the time that the dataset is uploaded. This will mean that the user can view if the dataset has been updated recently/ the next time that the dataset should be updated. This will remove time inaccuracy errors, as before a user might see that a dataset was uploaded 9 months ago and then search for a more up-to-date version, however if the data is only updated annually then this would most recently updated record of this dataset
- Data accuracy – as the dataset has to go through a validation/ pass data standards, this will ensure data accuracy
- Update time – the user will have to spend an initial period of time getting used to the new standard that the datasets adhere to
- Data format – the datasets will be in the same format
- Data availability – the user would be able to see what datasets are available instead of searching for datasets which are not published.

5.6.3 Defra and DAs

- Reduced costs and more effective decision making as a result of having better access to a wider range of comparable data.
- Automated (and therefore far more cost effective) reporting to the EU via EEA's EIONET / Reportnet systems via SEIS / INSPIRE approaches such defined XML "data flows"
- Possibility to build new tools on top of this data platform that will allow more effective decision making and better use of public money to tackle air quality issues.
- Less painful compliance with INSPIRE, SEIS and data.gov.uk initiatives as getting started early on and engaging with the relevant stakeholders and starting to move systems in the right direction technically.

6 Roadmap and Proposed Transition

Getting started on the journey towards integrated air quality data and INSPIRE compliance will be a complex and challenging process due to the number of stakeholders, number of systems and detailed datasets involved. Therefore Defra and the Devolved Administrations need to choose an effective strategy.

During the development of the proposed data integration structure, further stakeholder engagement will be required during the early stages to ensure a high level of commitment and to ensure the end results is focused on user requirements. It is envisaged that this could take place with a series of user focus groups organised to coincide with other air quality events, to attract a wide and broad range of users and to keep costs at a minimum.

The UK Air Quality Archive is currently under review and any future work in this area ought to also take into account the findings of that review, which is being carried out by Aether and Air Quality Consultants on behalf of Defra and the Devolved Administrations.

6.1 Proposed Strategy / Implementation Tasks

We believe one of the most effective strategies for Defra and the Devolved Administrations to allow for an integrated system and the improved resulting tools would be to:

1. Document current process and define requirements
2. Create a comprehensive catalogue of air quality data sets
3. Assess datasets for INSPIRE compliance
4. Refine and define UK data formats, metadata standards and overall architecture
5. Encourage data providers and other relevant stakeholder to start using and implementing the integration policy
6. Proceed with full INSPIRE compliance
7. Improve and develop tools

6.1.1 Document current process and define requirements

This should address the formats, metadata and uses of data. Each dataset provider needs to document their data flows and provide a data dictionary describing the data that is held / transferred publically. Any existing INSPIRE compliant datasets or mature robust datasets should be fed into stage 3.

6.1.2 Create a comprehensive catalogue of air quality data sets

By producing this catalogue, it will more straight forward for data users to find the available data sets; they could be put on data.gov.uk where applicable. Establish relationship with data.gov.uk and manage this via a Wiki or other collaborative Web approach.

6.1.3 Assess datasets for INSPIRE compliance

Evaluate which datasets are likely to fall under INSPIRE (and which Annex they will apply to) and engage with INSPIRE working groups (through the UK Location Programme) defining the standards to ensure the UK's view is represented and the future formats are manageable for the UK air quality data. Any existing specifications should be submitted and any gaps will illuminate the focus for step 4. Take work from stage 1 and feed into INSPIRE team. Keep abreast of relevant SEIS and INSPIRE developments.

6.1.4 Define UK data formats, metadata standards and overall architecture

As a parallel activity to 3 (utilising available INSPIRE standards) refine and define UK data formats, metadata standards and overall architecture that data providers & stakeholders need to comply with as part of Defra Integrated Air Quality data policy. These must build on existing work done to date on INSPIRE compliance and these standards can then be referenced in future procurement exercises. Part of this policy should outline how to handle non air quality data sets in a standard way (such as Health, Weather and Transport data).

Note that stages 3 and 4 will start the process of implementing the integrated system independent to INSPIRE, but in a future proof INSPIRE and SEIS compatible manner. An additional recommended step is to create a steering group on UK air quality data standards – to address implementation issues, build on existing work on INSPIRE compliance and ensure all stakeholders are engaged and involved throughout the process.

6.1.5 Encourage stakeholders to start implementing the integration policy

Encourage (or where possible enforce through procurement policy) data providers and other relevant stakeholder to start using and implementing the integration policy – the formats and technologies that support step 4 (including adopting outcomes of current INSPIRE compliant projects such as CAFE). This will make it easier to prepare for the future changes whilst reaping more immediate benefits that data standardisation will bring. Once the integration policy is mature enough dataset providers can start to implement compliance in their systems and datasets. This will involve:

- Adopting outcomes of current INSPIRE compliant projects such as CAFE
- Use of XML / RDF for data exchange
- Use of ISO and OGC standards referenced by SEIS and INSPIRE to make it easier to prepare for the future changes whilst reaping more immediate benefits that data standardisation will bring.
- Allocate budget for this work using the business case of more effective data supporting policy and tactical decision making, future savings and easier / less expensive INSPIRE compliance in the future.

Within the proposed solution there are two main options for data providers to choose between – either reengineering and reformatting all data from scratch, or applying an Export, Transform and Load process to automatically reformat all existing data on an ongoing basis. ETL tools are readily available for purchase and this data transformation approach could be a short-medium term solution, which could be applied to certain datasets to ease certain reporting requirements.

6.1.6 Proceed with full INSPIRE compliance

Once SEIS and INSPIRE implementation details are known and agreed proceed with full (or as near to full as practically possible) SEIS and INSPIRE compliance by revising the integration standards to align with these. By this point it should largely be a case of revising metadata and some services as the underlying technologies for data capture and transfer should be SEIS and INSPIRE compliant anyway.

6.1.7 Improving and developing tools

Once UK integration standards have been adopted (or SEIS / INSPIRE compliance has been added to the Air Quality Archive as an interim measure) new tools can be built to take advantage of the compatibility of the datasets. This is of course one of the key drivers for building the platform – to allow for new tools to be built. If there is a demand for new tools to be built before the UK integration standards are mature tool developers should consider how easy it will be for them to adapt their tools for SEIS/INSPIRE type compliance in the future.

6.2 Costs

It was not the intention of this scoping study to provide a detailed breakdown of the costs involved in the implementation of the integration system. However, through discussion with the high priority data providers we are able to provide some comment about the associated costs:

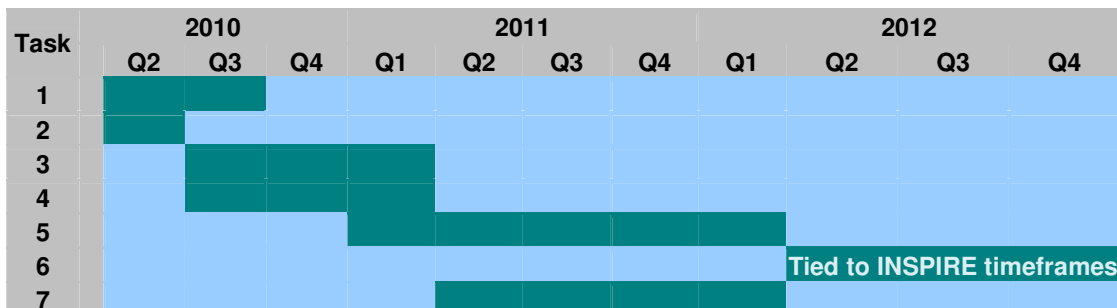
There are costs associated with:

- The creation of a detailed service architecture, development of data standards and other over-reaching activities which will impact all datasets, in the range of £60K to £100K depending on scope
- The creation of good metadata, which is vital to make the data useful, is likely to be a major cost, in the range of £120K to £200K. It is advised that this is undertaken before any modifications occur on individual datasets.
- The changes made to individual datasets. Costs per dataset have been illustrated in Section 5.1 as between £25K and £75K per dataset in addition to the costs that have occurred due to creation of good metadata.
- Implementation of the web services themselves will also be a significant cost. This is dependent on how far standard tools can be used to meet the specification, and the level of appropriate skills the data providers have to implement them.

6.3 Gantt Chart

The below shows the anticipated timeline for the implementation based on current information on INSPIRE timescales and integrated system requirements.

Figure 6.1 Gantt Chart showing anticipated timeline



1. Document current process and define requirements
2. Create a comprehensive catalogue of air quality data sets
3. Evaluate which datasets are likely to fall under INSPIRE
4. Define UK data formats, metadata standards and overall architecture
5. Encourage stakeholders to start implementing the integration policy
6. Proceed with full INSPIRE compliance
7. Improve and develop tools

6.4 Quick Wins

Maximising the use of the datasets and adding value for the data users is a key driver for this scoping study, equal to the requirement to start thinking about INSPIRE compliance and other regulatory requirements. It would be possible to start to add value to the data and make them more useable in the short term, without the bottom-up standardisation approach which has been outlined above.

A number of 'quick wins' have been identified by this scoping study, in particular during the final stakeholder workshop. These have been discussed throughout this report and are listed below. They include options which would be quick to implement, those which could be implemented at low cost and those which would be particularly good value (relatively low cost but high priority).

- Using this scoping study as a starting point, catalogue all useful air quality and non-air quality datasets and metadata, including those, but not exclusively those owned by Defra and the Devolved Administrations. Create appropriate web pages to list and provide links to these datasets. We anticipate the cost for this to be in the region of £25k.
- Integration of the continuous air quality data is a prime candidate for early integration. The data is relatively simple, the measurements are important and we have a lot of them. One option for this integration is to make use of RDFa – this is where RDF (a variant of XML) is added to existing HTML pages designed for human consumption (invisibly to human users) that adds support for use of the data systematically. We anticipate the cost for this to be in the region of £50k per dataset.
- Some data providers (KCL and BADC) hold local Met data. This could also be integrated with relatively little effort but with a great benefit (access to Met data has been identified by data users as a high priority). In particular, KCL have about 20 met sites in London and the South East whose data could be ready for integration in the short term. We anticipate the cost for this to be less than £100k.
- The collection and integration of metadata is viewed as a very high priority. Making these data available and easily accessible online is the first step to provide more meaning to the data. Standardising the format of this metadata and ensuring its completeness is important, but a more difficult and longer term task. We anticipate the cost for this to be in the region of £120k.
- If DfT can provide transport data in a consistent format these could be used to plug into LAQM tools. Further investigation is required to determine how DfT data could be readily transformed and automatically applied to currently available tools, but this is not expected to be a difficult nor expensive task. We anticipate the cost for this to be in the region of £25k.
- Identify reporting processes which can be readily simplified and automated and develop simple tools to do this.

All costs estimated in this section are provided as a guide only and full quotations from contractors will differ.

6.5 Alternative Strategy

One alternative approach would be to wait until the full details of INSPIRE (including Annex II) are known and defined before proceeding with implementation. This however is a risky approach that does nothing to mitigate the current issues around disparate datasets (not to mention preventing the full potential and future compatibility of any new tools developed in the meantime). It forces a much shorter UK implementation timescale for SEIS and INSPIRE compliance which will be more expensive, more error prone (due to less time for testing cycles) and riskier as a result.

This option was discussed at the final stakeholder workshop but not supported.

Appendices

Appendix 1: Regulatory Drivers

Appendix 2: Workshop 1 Minutes

Appendix 3: Data Users Survey

Appendix 4: Data Providers Survey

Appendix 5: Workshop 2 Minutes

Appendix 6: Dataset Summary and Scoring Criteria

Appendix 7: INSPIRE, SEIS and Data.gov.uk

Appendix 1

Regulatory Drivers

EU Directive on the Reuse of Public Sector Information.

Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information provides a common legislative framework and guidelines for public sector organisations to ensure the availability of data.

In Article 3 of the Directive, the General Principle is stated,
“Member States shall ensure that, where the re-use of documents held by public sector bodies is allowed, these documents shall be re-usable for commercial or non-commercial purposes in accordance with the conditions set out in Chapters III and IV. Where possible, documents shall be made available through electronic means”.

The premise of the Directive is that allowing and promoting the re-use of information will ensure fair competition, transparency and support the development of cross-border services and products based on public sector information. This includes all spatial environmental and air quality data covered by the INSPIRE Directive.

In May 2009 there was a review of the Directive⁹. It concluded that significant progress has been made in making public sector information non-exclusive and available for re-use, but that there are still barriers to success and access to some data is still restricted.

CAFE Directive

Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe sets out requirements for information and reporting, stating,

“It is necessary to adapt procedures for data provision, assessment and reporting of air quality to enable electronic means and the Internet to be used as the main tools to make information available, and so that such procedures are compatible with Directive 2007/2/EC of the European Parliament and the Council of 14 March 2007 establishing an infrastructure for spatial information in the European Community (INSPIRE)”

In Chapter V, Articles 26-28 it specifies air quality data, annual reports and relevant information that must be made available to members of the public and other stakeholders.

Currently data reporting to the European Commission is via more than one method, and there is no standard procedure across Member States or within the UK across different datasets, for reporting of levels exceeding limit values, target values, long-term objectives, information thresholds and alert thresholds.

Integration and harmonisation of the UK's datasets will allow easier and quicker interrogation of the data and a simplified process for reporting to members of the public, stakeholders and the Commission.

INSPIRE

Directive 2007/2/EC of the European Parliament and of the Council came into force on 14 March 2007, establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). This initiative creates the underlying rules for exchanging data and services across national boundaries in Europe. In Europe, national borders have, for many years, hindered the development of cross-boundary services in such diverse sectors as transport, planning, emergency services, natural resources, environmental monitoring, business enterprises and provision of utilities. There are a multitude of barriers, including the obvious language and regulations, to the less obvious availability and format of data, in particular those data that relate to location.

In these days of technological advancement it has become apparent that there is a requirement to, and we have the capability to, remove these barriers and reap the benefits, through the development and implementation of a Europe-wide infrastructure for spatial information. This would support the integration and harmonisation of huge quantities of spatial (location) data from multiple sources in each of the Member States into a single framework.

⁹ [http://ec.europa.eu/information_society/policy/psi/docs/pdfs/swd_070509/re-usepsi_sec\(2009\).pdf](http://ec.europa.eu/information_society/policy/psi/docs/pdfs/swd_070509/re-usepsi_sec(2009).pdf)

The principles of INSPIRE are:

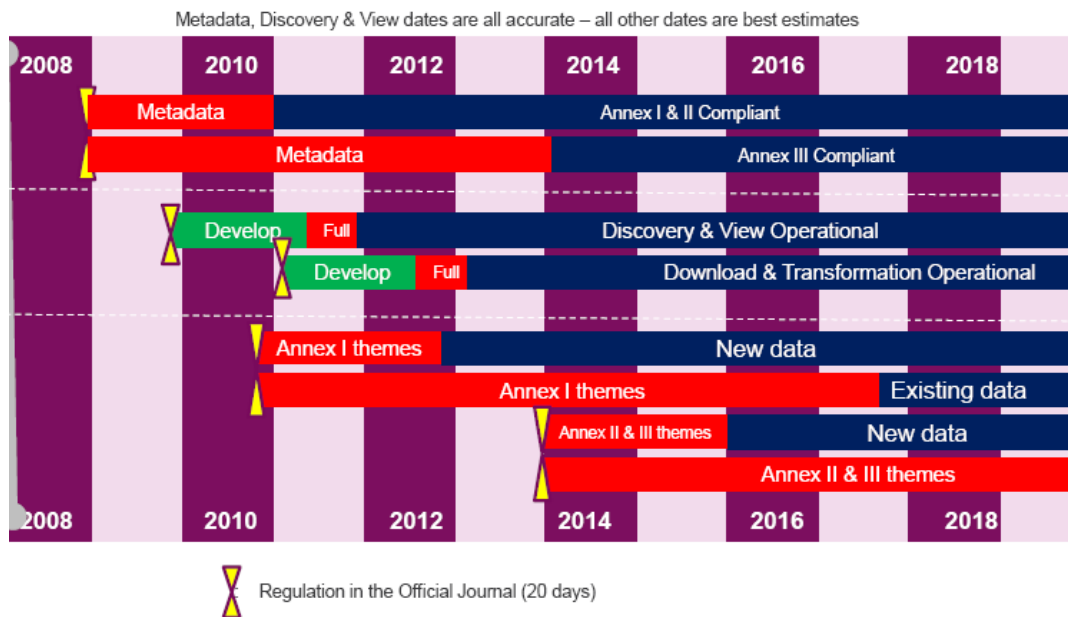
- Data should be collected once only and shared between all levels of government and all stakeholders
- It should be possible to combine spatial data from different sources
- It should be easy to discover which spatial data are available, to evaluate if they are fit for purpose and to know what conditions apply for their use.

The Directive includes metadata regulation, data specification regulation, network service regulation, data & service sharing regulation and monitoring & reporting regulation. The Member States are expected to follow these regulations when they are published.

The INSPIRE initiative is currently in its final phase – preparing for implementation. The implementation of the services will occur in the following order: metadata, discovery and view data services, download and transformation services, a map feature service and interpretative services.

The Timescale for INSPIRE⁹

Figure A1.1 INSPIRE Timescale



Different time schedules are linked to the data in the three annexes I, II and III. Air quality monitoring stations fall within the ‘Environmental monitoring facilities’ in Annex III, alongside meteorological stations. The scope of this theme is defined¹⁰ as follows:

“Environmental monitoring facilities are facilities for observations and measurements of emissions, status and effects of environmental media (e.g. air, forest, marine water) and/or other environmental aspects (e.g. biodiversity, human health). The concept of monitoring may relate to systematic and hierarchical structures, including monitoring networks, monitoring stations, monitoring site and subsites. The monitoring sites may be permanently located at a site or can be temporal, only used for a certain time. Continuous moving monitoring facilities, e.g. on ships, may be a kind of monitoring facility. Monitoring sites in the form of locations and areas can be reported as georeferenced points, lines and polygons. In cases where data are classified or confidential, aggregation to grids may be a possibility.”

⁹ http://www.aqi.org.uk/SITE/UPLOAD/DOCUMENT/Event_Presentations/0091118UKLP/KeithMurray1.pdf (Slide 14)

¹⁰ Drafting Team “Data Specifications” – deliverable D2.3: Definition of Annex Themes and Scope v3.0, March 2008

Air quality data are also related to themes in other Annexes as explained in Appendix 7.

UK Location Strategy

The UK Location Programme is the vehicle through which the UK will implement the UK Government's Location Strategy¹¹, and the INSPIRE Directive.

The Location Programme will comply with all other relevant regulations discussed here and provides an immediate driver for this scoping study and the potential integration of the UK's air quality data.

It has been designed, not only to meet the regulatory requirements of European Directives and initiatives, but also to reduce costs of delivery by cutting duplication of effort, and to improve and add value to our existing datasets. The Location Programme encompasses a diverse range of sectors, of which environmental monitoring is one. The integration of air quality datasets is a small part of this large programme and as such care must be taken to ensure that the integration is undertaken within the guidelines of, and in collaboration with the UK Location Programme.

The INSPIRE team at Defra have been consulted during this scoping study and if the proposed, or indeed any other, air quality data infrastructure is implemented the INSPIRE team should continue to be involved to ensure that compliance is fully and correctly considered.

Shared Environmental Information System

The Shared Environmental Information System (SEIS) is a joint project of the European Environment Agency and the European Commission. The objective is to develop a system that would allow Member States to share all environmental data and related information electronically, by simplifying data flows and employing modern technologies. The SEIS relies on the compatibility of data from all Member States and requires collaboration between Member State Governments and data management contractors to provide suitably managed and integrated datasets.

The Commission's Expert Group on Data Exchange has recently prepared a working document to identify the requirements of a reporting system for ambient air quality legislation in Europe, so the Commission can prepare Implementing Provisions on Reporting. It is expected that these measures will be adopted during 2010. The Implementing Provisions will specify how the reporting requirements set out in the CAFE (2008/50/EC) Directive and INSPIRE (2007/2/EC) Directive will be met by the Member States. Development of this electronic streamlined data exchange and reporting system will also contribute to the development of SEIS. The main elements of the Implementing Provisions, as described in the working document, shall be:

- Specification of the information to be reported
- Information flow requirements (deadlines/periodicity, reporting scheme etc.)
- Common data format and metadata description
- Description of tools for checking the format, data/metadata consistency and integrity
- Description of rules/tools for merging, aggregation and rendering of the data.

The recommendations outlined in this report have been developed to compliment these main elements, however, if the UK's air quality data is integrated in the future care must be taken to ensure that the integration is compatible with the Implementing Provisions guidelines.

SENSE

The European Environment Agency are leading a project for Shared European and National State of the Environment (SENSE), which will be part of the stepwise implementation of the SEIS. The UK is expected to participate in this project, which aims to *"implement a mechanism for SEIS compliant information sharing between national and European websites based on further enhancement of Reportnet, aiming at more timely and cost-efficient quality assured data flows"*¹² during 2010.

¹¹ Place matters: the Location Strategy for the United Kingdom, November 2008, Geographic Information Panel

¹² SENSE Project Plan draft, December 2009, European Environment Agency

Appendix 2

Workshop 1 Minutes

Data Management and Integration System Workshop

15 January 2010, Defra

Present

Andrew Monteith (AEA)
Anne Wagner (AEA)
Clare Bayley (Defra)
David Butterfield (NPL)
David Carslaw (KCL)
Emily Nicholl (Defra)
Gary Fuller (KCL)
Gwyn Jones (AEA)
Ian McCrae (TRL)
Iarla Kilbane-Dawe (AEA)
Neil Cape (CEH)
Ollie Cronk (AEA)
Paul Willis (AEA)
Rachel Yardley (AEA)
Ray Evans (BV)
Richard Maggs (BV)
Robert Hepburn (KCL)
Susannah Grice (AEA)
Tim Williamson (Defra, chair)

Context and drivers of the study

- Iarla outlined the drivers for a data management and integration system, explaining what data integration involves and highlighting the outcomes required from the workshop. Several examples of data management tools were demonstrated.
- The primary purpose of the proposed data management and integration system is to ensure that data produced with public money should be made available to the public.
- The aim of this scoping project is to deliver a feasibility report to Defra and the DAs by the end of March 2010, with options for their next steps. It is important to note that this is not a data-mining project; it is a data integration project. A large part of the study will be to investigate what data is currently available and what needs to be done to the data in order for it to be able to be integrated.

Exercise 1: What data do we have and how is it used currently?

- Paul gave an overview of the purpose of this exercise, and short presentations of data were given
 - Paul Willis – Archive
 - Rachel Yardley – AURN data
 - Gwyn Jones – LAQM data
 - Anne Wagner – NAEI data
 - David Butterfield – Metals, black carbon, hydrocarbons and particle speciation data
 - Richard Maggs – AURN data
 - Neil Cape – EMEP, UKEAP and rural heavy metals data
 - Susannah Grice – PCM data
 - Gary Fuller – LAQN, LAEI, met and traffic data
 - Ian McCrae – Highways Agency sites, traffic data, Local Authority data
- The following points were made during the discussed that followed the presentations:
 - The Air Quality Archive is an example of best practice in sharing air quality data. It is updated 15 minutes after every hour and has a standard data format so people can write their own scripts.
 - The frequency of the data updates will vary by data set, and will be dependent on ratification and processing procedures and other delays.
 - It is important to minimise the data set size, as in a previous example the metadata on top of the standard data increased the size of the data by a factor of over 100.
 - The data sets will be of various temporal lengths - some of the data has been held for decades and some of it has been only a limited history. The quality and integrity of the datasets is also likely to vary over time.
 - Local Authorities (LA) use their own systems to produce the reports that they need. Ideally the proposed integration system should enable Local Authorities to add their own data to the system.
 - CEH is currently developing an Environmental Informatics Data Centre, which is managed by Gwyn Reiss.
- There was some discussion around the integration of non-AQ data:
 - The TRADS traffic database is freely available, otherwise traffic data may be provided by the Highways Agency and/or TfL.
 - Satellite data from GMES could be considered
 - Land use data could be considered.

Exercise 2: Discussion of proposal and functionality

- David Carslaw presented the Openair tools
- The workshop identified several potentially useful tools which could be developed to work in association with the proposed data integration system:
 - Health impact tool
 - Data ratification tool
 - Asset management/investment programme tool to allow Defra to manage their investments and to aid forward planning
 - Air mass back trajectory tools
 - Tools to quantify the benefits of certain actions (see the Dutch and EPA tools)/testing different scenarios

- Presentational tools (mapping and graphing) on the local and national scale
 - Applications for ecosystems, greenhouse gases, agriculture
 - Tools for Local Authorities, to reduce the burden of data gathering and Review and Assessment and to improve action planning tools, local and generic data tools.
- The following barriers, suggestions and considerations were discussed:
- It is important that the proposed system doesn't just make copies of datasets – having multiple copies should be avoided
 - The proposed integration system could be designed to have a geographical interface
 - All user interfaces to the proposed system should be simple and easy to use
 - The scoping study ought to consider permission required by non-Defra/DA data sets and also consider licensing issues
 - Met data and traffic data may be difficult to access.
 - It is important that we understand the feasibility of integrating Defra AQ data sets before we consider non-AQ data integration.
 - How will we deal with text data that is not of specific controlled format?
 - Is data ratification done consistently?
 - Met office data is expensive – can we use WRF instead to provide forecast Met data?
 - The proposed data integration system should include a list of all available data, plus metadata, in a data dictionary
 - MOLES (Metadata Objects for Linking the Environmental Sciences) is a good example of an information model that provides a framework to support metadata, from NERC and BADC.
 - Consider OLAP (Online analytical processing) cube as a data structure that enables rapid data analysis.
 - What data cleaning/sorting/formatting would be required to get the data ready for integration?

Appendix 3

Data users' survey

Air Quality Data Management and Integration System Scoping Study

Data Users Survey

Defra has commissioned a scoping study to consider the viability of an Air Quality Data Management and Integration System. The key objective of the study is to investigate and define what needs to be done to create a framework that ensures the accessibility and re-usability of Defra's key air quality data investments.

The purpose of the system would be to open up and simplify data use for all stakeholders and to provide online compliance and analysis tools. Such a system will have significant benefits for Defra and the Devolved Administrations, Local Authorities and the wider Air Quality community.

As a user of air quality data, you are invited to help shape the future data management system in the UK.

This study intends to draw out the changes which need to be implemented to allow UK air quality data to be integrated, the difficulties currently faced by the users of air quality data, and potential tools to aid air quality data analysis in the future.

Please send completed questionnaires and direct any questions to rachel.yardley@aeat.co.uk.

In some cases it may be more appropriate and easier to discuss these questions by telephone. This is likely to take about 15 minutes. If you are happy to be contacted during the next few days please contact rachel.yardley@aeat.co.uk to suggest a date and time that is suitable for you.

General Questions

1. Name:
2. Company:
3. AQ Involvement: Data provider Data user Other
4. Telephone number:
5. Email Address:
6. Would you be prepared to discuss your response in more detail if required?

Air Quality Data Users

1. What data do you use most regularly to inform air quality research, analysis and decision making (e.g. local or national monitoring data, emissions inventories, maps and forecasts)?
2. What air quality and other related data do you analyse and/or report on and why?
3. Where do you currently get these data from (websites, databases, internal, external)?
4. Are there any data that you would like access to but don't have?
5. How much time do you currently spend gathering, formatting and analysing air quality and related data each year?
6. How do you currently rate access to UK air quality and related data and air quality analysis tools?

Excellent Good Adequate Inadequate Awful

Please indicate whether you agree or disagree with the following statement. Please make additional comments if you wish:

7. I have access to all the **air quality** data I need

Agree No opinion Disagree

Comment:

8. I have access to all of the related **non air quality data** I need (for example, population data, health statistics, traffic land use, meteorological data etc.)

Agree No opinion Disagree

Comment:

9. It is quick and easy to manipulate and analyse the data that I require

Agree No opinion Disagree

Comment:

10. I would welcome a single air quality portal with improvements to the availability, format and integrity of these data

Agree No opinion Disagree

Comment:

11. I would welcome basic or complex online tools to help me analyse these data

Agree No opinion Disagree

Comment:

12. I would welcome basic or complex online tools to help me present (tables, graphs, maps) these data

Agree No opinion Disagree

Comment:

13. I would welcome basic or complex online tools to help me make decisions regarding air quality in the UK

- Agree No opinion Disagree

Comment:

14. I would welcome basic or complex online tools to help action planning

- Agree No opinion Disagree

Comment:

15. I would welcome more centralised, automated and uniform tools and methods for data analysis and reporting

- Agree No opinion Disagree

Comment:

16. I have experienced problems with current spreadsheet-based data & tools (e.g. out-of-date, version control etc.)

- Agree No opinion Disagree

Comment:

17. What are your primary problems or concerns with accessing, manipulating, analysing and presenting air quality data?

- Limited availability Technically difficult Time consuming
 I don't have the right data analysis tools I have no problems

18. Please describe any specific difficulties that you have:

19. If you have any suggestions for air quality tools which would be particularly useful to you, please describe these below. For example:

- *Local Air Quality Management or national air quality compliance tools*
- *Presentational tools (e.g. maps, graphs, data tables)*
- *Practical tools (e.g. to ratify data or manage assets)*
- *Predictive tools*
- *Action planning/impact analysis tools*
- *Emissions scenario testing tools*
- *Analysis tools (e.g. to compare pollutant emissions with air quality, to compare pollutant concentrations with health statistics)*

For reference, examples of existing tools can be found at <http://www.airquality.co.uk/laqm/tools.php> (mainly spreadsheet based), or the Netherlands online Clean Air Policy Tool (www.saneringstool.nl/saneringstool_ENG.html)

Appendix 4

Data providers' survey

Air Quality Data Management and Integration System Scoping Study Data Providers Survey

Defra has commissioned a scoping study to consider the viability of an Air Quality Data Management and Integration System. The key objective of the study is to investigate and define what needs to be done to create a framework that ensures the accessibility and re-usability of Defra's key air quality data investments.

The purpose of the system would be to open up and simplify data use for all stakeholders and to provide online compliance and analysis tools. Such a system will have significant benefits for Defra and the Devolved Administrations, Local Authorities and the wider Air Quality community.

As a provider of air quality data, you are invited to help shape the future data management system in the UK. This study intends to draw out the changes which need to be implemented to allow UK air quality data to be integrated, the difficulties currently faced by the users of air quality data, and potential tools to aid air quality data analysis in the future.

Please send completed questionnaires and direct any questions to andrew.monteith@aeat.co.uk. In some cases it may be more appropriate and easier to discuss these questions by telephone. This is likely to take about 30 minutes. If you are happy to be contacted during the next few days please contact andrew.monteith@aeat.co.uk to suggest a date and time that is suitable for you. It may be necessary to involve your IT colleagues and if this is the case we can arrange teleconferencing facilities.

General Questions

1. Name:
2. Company:
3. AQ Involvement: Data provider Data user Other
4. Telephone number:
5. Email Address:
6. Would you be prepared to discuss your response in more detail if required?

Air Quality Data Providers

If you have any difficulty answering these questions please contact a member of your IT team, or contact Andrew Monteith on 0870 190 6596.

1. What is the name of the dataset?
2. Who are the data collected for and which contract are they collected for?
3. For what purpose are the data currently used (please include relevant legislation and reporting requirements)?
4. What data fields are included in the dataset?
5. Please describe the metadata which are available for this data set, and the format of the metadata
6. What is the period that the data are collected over (archive length)?
7. Do you supply the same approach/ methodology across the whole time series?
8. What is the format of the dataset? (e.g. Industry Standard / XML or Proprietary)
9. How often are the data updated?
10. How often is the data format edited?
 Weekly Monthly Annually Never
11. How often will the data format change in the future?
 Weekly Monthly Annually Never
12. What is the geographical coverage of the data collected?
 EU National Local
13. Please describe the spatial resolution of the data (i.e. grid-based, point source):
14. Please describe the temporal resolution of the data (i.e. hourly, daily, annual mean):
15. What are the external datasets currently used or assimilated?
16. What are the methods by which the data is accessed?
 Individual copy On a shared drive Distribution list/ Email
 Downloaded Online
17. What is the approximate size of the dataset (in GB)?
18. Does the dataset use web service standards? If yes, state which standards are used:
19. How many servers are used to store the data?
20. Are the data unstructured, if no can you supply a relational diagram if applicable?
21. How many total users are there for the dataset?
22. How many users are there during a typical peak period?
23. Is security in place to limit the data that each user group can view?
24. Data security methods:
 SSL (Secure Sockets Layer) Secure data center ISO27001 compliant
 Other, please give details:
25. Are there any Data Ownership, Data Protection and IP Issues – e.g. are the data in the public domain / Crown Copyright or restricted Intellectual Property?

26. Please describe the quality assurance and quality control procedures which are applied to the dataset:
27. What database system is used to store the data?
28. What is the database model?
29. How are queries to the databases structured to extract information?
30. How are users asked to structure their queries and using what fields or keys (i.e. by time, date, species, location, etc)?
31. How are queries sent (through a manual interface web page or through a web service that allows automated access)?
32. If there is anything else you would like to add about the data that are collected please state it here:

Please send an example copy of your dataset to Andrew.Monteith@aeat.co.uk

Thank you for taking part in the UK's air quality data integration study. The results of this exercise and recommendations will be available to you by the end of March 2010.

Appendix 5

Workshop 2 Minutes

Air Quality Data Management and Integration System Scoping Study
Defra, Ergon House
12th March 2010, 1.30pm

Defra and the Devolved administrations

Tim Williamson (Defra) Chair
Clare Bayley (Defra)
Ross Hunter (Wales)

Apologies

Susannah Grice (AEA - PCM)
Neil Cape (CEH)
Ray Evans (BV)
Gary Fuller (KCL)
David Butterfield (NPL)
Bryan Lawrence (BADDC)
David Carslaw (KCL)
Stephen Kerr (Northern Ireland) tried to dial in but technical difficulties prevented this

Data Providers and project team

Robert Hepburn (KCL)
Richard Maggs (BV)
David Leaver (CEH)
Paul Willis (AEA)
Iarla Kilbane-Dawe (AEA)
Julius Mattai (AEA)
Andrew Monteith (AEA IT solutions)
Ollie Cronk (AEA IT solutions)
Rachel Yardley (AEA - AURN)
Andrew Kent (AEA – PCM)
Anne Wagner (AEA - NAEI)
Steve Moorcroft (AQC)
Ag Stephens (BADDC)
Sue Grimmond (KCL)
Gwyn Jones (AEA)
Xingyu Xiao (AEA)

The purpose of the workshop was to present the findings of the scoping study, outline a proposed solution and discuss the benefits and issues surrounding this. A draft report was distributed to all attendees before the workshop and the presentation slides are circulated with these minutes. All suggestions and comments discussed during the workshop will be considered in the final report.

Presentation slides

- Welcome and Introductions led by Tim Williamson
- Workshop 1 recap and Activity since Workshop 1 by Paul Willis
- Highlights from stakeholder survey (users) by Rachel Yardley
- Highlights and Headlines from stakeholder survey (data providers), INSPIRE requirements, aims and objectives, Recommendations for Integration and Roadmap for future by Ollie Cronk

Stakeholder comments and feedback

- The authors should ensure that the dataset scoring is not biased towards AEA due to a lack of understanding of the other datasets. The scoring needs to be reviewed.
- The metadata referred to in the INSPIRE timeline are not about monitoring stations but rather IT/high level metadata about language, formats etc. The detailed metadata will be included in the CAFE Directive Implementing Provisions and these details can then be captured in INSPIRE compliant reporting.
- CEH data ought to be better represented. This had not been included due to late receipt of the relevant data but will be added to the final report.
- The group agreed that the 'current architecture' diagram is representative of the current situation, and Robert Hepburn confirmed that much of KCL data are in a similar structure. If anything, the diagram underplays the complexity of the current situation due to the very limited number of datasets represented.
- Within the proposed solution there are two main options for data providers to choose between – either reengineering and reformatting all data from scratch, or applying an Export, Transform and Load process to automatically reformat all existing data on an ongoing basis. ETL tools are readily available for purchase. This data transformation approach could be a short-medium term solution, or 'quick win' which could be applied to certain datasets to ease certain reporting requirements
- The proposed new approach will allow much simpler EIONET reporting.
- Robert Hepburn stated that KCL already use a Standardised data Update Service layer so care needs to be taken that work is not duplicated. However, the Data Access Service Layer is a very good idea.

- The group agreed that while our community is very good at collecting data, we are not so good at using it for making decisions. This should be highlighted and more focus should be on the use of tools in the report.
- If DfT can provide transport data in a consistent format these could be used to plug into LAQM tools. The report should highlight transport data as a priority for integration.
- There are ongoing issues with data access and cost of data from the Met Office. BADCs have a lot of Met Office datasets and they are allowed to make them available to certain other users. The proposed new structure would contain an element for controlling user access to the datasets, so it may be possible to allow certain users access to Met Office data in this way. The report should highlight Met data as a priority for integration.
- During the development of the data integration structure, further stakeholder engagement will be required
- The use of, or conformation with the new structure should be stipulated by the Defra and DA procurement process
- Tim Williamson clarified that the main purpose of the proposed integration would be to make maximum use of the data we have, not to be INSPIRE compliant (though this would be an obvious benefit). Other attendees were keener that this integration study should be seen as an early first step towards INSPIRE compliance.
- The data catalogue/dictionary has full support of the group and is an obvious first step in the process. It would also be useful to publish the annual data processing cycle alongside this catalogue. This should be highlighted in the report as a high priority.
- Any tools that are developed as a result of the integration should be easy to use for non-IT experts, in particular for LA environmental health officers.
- The focus of the scoping study has been the nature of the current data and the platform for integration, rather than the possibilities for new tools, which were a secondary and lesser part of the specification. Although the proposed robust data platform is an excellent starting point, future tool requirements may highlight other data that need to be captured in the integration system.
- The report should pick out quick wins for early, low risk, low cost implementation. For example, access of DfT data, simplifying EIONET reporting, developing a data catalogue. An estimation of costs to implement would be very useful.
- The report should more clearly prioritise the recommendations and next steps, for example, tool development, focus groups.
- A concern was raised that the proposed IT solution should not restrict future options or development of the system, nor tie us in to using specific software.
- Ollie Cronk warned that there are difficulties associated with adding web services to Excel and Access datasets – this is likely to be more of a limitation/issue than whether the underlying database is proprietary (as long as it uses or supports standards (such as SQL/XML) and can be scaled and securely web enabled).
- The report should highlight that the system and resulting tools must be user-driven.
- Ross Hunter suggested a tool that would be useful for Local Authorities would be one which allowed the user to select a certain area and the tool returns all data known about that area (both air quality and non-air quality)
- The datasets considered in the scoping study were selected by Defra to be a representative sample and it is not necessary to consider any others at this stage.

Appendix 6

Dataset Summary and Scoring Criteria

Scoring Matrix:

The scoring matrix is a method by which to rank the top priority datasets; NAEI, PCM, AURN, LAQN, NPL (an average of the Hydrocarbon, UK Heavy Metals, and UK Black Carbon and Black Smoke networks) and CEH (an average of the Rural Heavy Metals Monitoring and EMEP network) across a range of different criteria, comparing them with an ideal dataset. This evaluation has assessed datasets against a new set of criteria that previously the datasets have not had to comply with. Therefore lower scores do not mean that the datasets are not fit for their current purpose, only that they will require more transformation to integrate them. Below is a description of the scoring for each of the criterion.

Data Searchability - How easily searchable and understandable the dataset is to the user.

0 – Data cannot be found, i.e. there is no search engine.

5 – The dataset can be found after typing in key words related to that dataset

Data Timeline - The dataset contains up-to-date data.

0 – Data are rarely updated if at all after it has original been published.

5 – Data are frequently updated, and there is an indication of when the last update took place/ when future updates will take place.

Data Downloadability - How easy the data are to download and the usability of the data for analysis.

0 – The data cannot be downloaded.

5 - The entire dataset can be downloaded and used for other applications for analysis.

Data Historical Trending - Does the dataset contain records from past years?

0 – Contains no previous data

5 – Contains over 20 years worth of available data

Data Comprehensiveness - The detail and the area that the data are collected over.

0 – Data are collected disparately.

5 – Data are collected at regular intervals, and with in-depth detail.

Data Accuracy - The correctness of the data and the methods that are in-process to ensure as little data-inaccuracy occurs as possible, i.e. through programmatic components which ensure data types.

0 – Data are inaccurate, out of date and are presented in a variety of different forms.

5 – The data are kept regularly updated, and there are automated checks to ensure that the data recorded are reasonable and acceptable.

Data Consistency – Are the data kept in the same format, and the same methodology used to capture the data.

0 – Data are kept in a variety of formats, and different methods have been used to record the data.

5 – The data are recorded through a defined standard, designed to reduce erroneous data, the same approach has been used to collect the data and if it has altered, then previous data have been modified accordingly.

Metadata – The amount of details that are given regarding the dataset; data about the data.

0 – No information about the data is given.

5 – There is a detailed structured list regarding the details of the dataset, which is stored in a human-readable format such as XML which is compliant with INSPIRE regulation.

Data Format and Standards - Are the data stored in a format which enables to easily integrate the data into future analysis?

0 – The data are stored in a format which hinders future integration.

5 – The data are stored in XML and are compliant with INSPIRE regulations (Please note that no dataset can score a 5 as to date the INSPIRE regulations have not been produced)

INSPIRE – Does the data comply with INSPIRE standards?

0 – The data are not in any shape or form compliant with INSPIRE standards, or other EU standards.

5 – The data are fully compliant with INSPIRE regulations. (Please note that no dataset can score a 5 as to date the INSPIRE regulations have not been produced)

Table A6.1 NAEI, PCM, AURN and LAQN Scoring Table:

Criteria	Ranking						
	NAEI	PCM	AURN	LAQN - Continuous AQ_DG	LAQN - MET	LAQN - Metals	LAQN - Traffic
Data Purpose	The purpose that the data are collected is to comply with EU and internal reporting requirements (GHG Monitoring Mechanism [UNFCCC], NEC Directive and LRTAP Convention [UNECE]).	Data compiled specifically for a) reporting under EU ambient air quality directives and b) air quality policy developments.	Relevant regulations: EC Directive on Cleaner Air For Europe UK Air Quality Strategy Objectives Data are used routinely for the following purposes: Reporting via the Questionnaire and Data Exchange Module Validating pollution climate models Used by AQEG and research groups Used by Local Authorities for LAQM Air quality forecasting and public air quality alerts	Affiliated LAQN data – reporting to the EC under Air Quality Directive. Volatile Correction Method (VCM) corrected PM ₁₀ TEOM concentrations from AURN sites - reporting to the EC under Air Quality Directive	The data stored are very specific as they are designed for research purposes	Reporting to the EC under Air Quality Directive (lead) and 4th Daughter Directive (arsenic, cadmium, mercury, nickel).	The data collected contains Site, DateTime, Road Lane, Vehicle class group (by axel length), speed group, number of vehicles

Criteria	Ranking						
	NAEI	PCM	AURN	LAQN - Continuous AQ_DG	LAQN - MET	LAQN - Metals	LAQN - Traffic
Data Searchability	2 - it is very easy to find the general data that you are looking for, as the user selects the pollutant they wish and then they are provided a summary table of that pollutant since 1970. However it is hard to search for a specific data set, as there is no search site specific search engine. Only the last previous years dataset is searchable, if you wish previous years only the year total is shown.	1 - The majority of the data cannot be found on the AQ archive. There is no search engine, the navigation is not immediately intuitive, however once the user gets accustomed to the site, it is very easy to navigate and find the desired dataset. Some of data have to be requested	1 - The majority of the data can be found on the AQ Archive. There is no search engine, the navigation is not immediately intuitive, however once the user gets accustomed to the site, it is very easy to navigate and find the desired dataset.	1- the data are searchable via a manual web page interface, and can be queried by site, species, date, time etc.	1 - the data are searchable via manual web page interface, and can be queried by site, species, date, time etc.	0 - The data are currently available via the website (there are plans to do this). So the data is not easily searched as there is no method at the moment in which to do this. The user would have to request the data	0 – The data are not currently available via the website, so it cannot be searched via a search engine. Users cannot query the database to extract information
Data Timeline	2 - The data are uploaded in a annually cycle.	2 – All the metrics covered by the PCM model are for an entire calendar year. Hence the smallest possible time step for providing new base year maps is every calendar year.	3	5 - The data are updated hourly, with 15 minute readings	5- The data are updated hourly, with 15 minute readings	4	3 – Monthly updates.

Criteria	Ranking						
	NAEI	PCM	AURN	LAQN - Continuous AQ_DG	LAQN - MET	LAQN - Metals	LAQN - Traffic
Data Downloadability	1- Only the last year's datasets can be downloaded to show the breakdown by region. If you wish to have previous data you can only download the year breakdown.	3- The majority of the data cannot be easily downloaded; however LAQM background maps can be accessed by anyone on the AQ archive. Individual request for maps are dealt on a case by case basis	2- The majority of the data can be downloaded from the AQ Archive	2 - the user can download the results of the query that they form, the data is in the public domain	2 - the user can download the results of the query that they form, the data is public domain	1 – As the data would have to be requested by a member of the public, the data is not easily downloadable.	1
Data Historical Trending's	5 - the database holds record for 1970-2008, and then predicted forecast for 2010,2015,2020	3 - Reports can be produced since 2001 Graphs are produced for the base year , which is the previous year. For selected pollutants projected data for 2010, 2015 and 2020 are also available	5 - The database holds data dating back to 1970.	2 - 1996 to the present (on-going)	2 - 1997 - to present	2 - 1997 - to present	2 - 1997 - to present
Data Comprehensiveness	4 - The data are collected over national geographical coverage. The spatial resolution is National data, point source data for	4 - 1x1 km GIS maps of background concentrations across the UK for NOX, NO ₂ , PM ₁₀ , PM _{2.5} , Roadside concentration	3 – point source	3 - the spatial resolution of the data is point measurements	3 - the spatial resolution of the data is point measurements	3	1

Criteria	Ranking						
	NAEI	PCM	AURN	LAQN - Continuous AQ_DG	LAQN - MET	LAQN - Metals	LAQN - Traffic
	power plants, spatial data for latest year on 1x1km grid	maps are available for major roads in urban areas					
Data Accuracy	3 - It is possible that previous data might need to be changed, i.e. due to a change in the method that data is collected. However the data need to go through a rigorous checking and sign off phase before they are published. Data also validated by cross checks with PCM data	3 - Independent verification of model results using non-AURN sites Routine independent checking procedures for modelling. Reality checks on data and comparison with previous years.	3 - Validation of data is carried out in accordance with the procedures set out in the QA/QC manual for the AURN	3- QA/QC undertaken by AEA. Volatile Correction Model (VCM) corrected PM ₁₀ TEOM concentrations from AURN sites – QA/QC of TEOM and FDMS measurements used in the correction undertaken by AEA	2 - Site inspections, automated and manual validation algorithms.	2 - Traceable flow checks on sampling. Analysis was undertaken using UKAS accredited ICP-MS procedures following microwave nitric acid digestion of the filter samples	2 - Site inspections, automated and manual validation algorithms.
Data Consistency	3 - Previous data are liable to change if the method of data collection is altered. However the datasets are structured in the same format	4	4 – the format of the data has not been changed	2 - Prior to 2004 in SE England TEOM data are multiplied by 1.3, after this date TEOM data are corrected using the VCM.	5 - the data contains the same approach/ methodology across the whole time series	4 - the data contains the same approach/ methodology across the whole time series, but it might need to be altered in the future	4

Criteria	Ranking						
	NAEI	PCM	AURN	LAQN - Continuous AQ_DG	LAQN - MET	LAQN - Metals	LAQN - Traffic
Data Processing	2 - There are a large number of worksheets which need to be manually edited after they have been submitted by the user.	2	1 - Data is uploaded through a multiple of different sources (FTP, email, manual entry, network monitoring) through a range of different mediums (.CSV, Excel, text)	4 - The data is automatically uploaded and processed.	1	1	1
Metadata	3 - Very detailed source specific data from various providers are supplied. There is a standard structure.	1 - There is a report which is updated yearly.	2 - The metadata for the contract are shown in the Site Information Archive (http://www.bv-aurndata.co.uk) and the asset register for equipment as provided to Defra (serial numbers, equipment type (manufacturer), age, etc.	3 - Location of site (grid ref and lat long), head height of measurement, species descriptions, all via website	3 - Location of site (grid ref and lat long), height of measurement, species descriptions, all via website	2 - Location of site (grid ref and lat long)	1 - Location of site, information about lanes and vehicle and speed groups
Data Format and Standards	2 - Access and XLS, The quality assurance and quality control procedures are in line with IPCC Guideline for Emission Inventories. The years datasets which are downloadable are well structured	2 - GIS background maps come as ESRI arc GIS grids GIS roadside maps come as ESRI arc GIS coverage's Exceedance statistics, source apportionments etc. available in spreadsheets LAQM background maps as .CSV files One server with daily backups.	2 - There is no standardised format for uploading the data, which results in long QA/QC process, and people having to manually edit the incoming data before it can be uploaded onto the system	2 - Uses SQL Server 2005, and the data is provided to external uses as .CSV 4 servers: 1 master database server, 1 live mirror, 1 web database server (updated hourly), 1 standby web server	2 - Uses SQL Server 2005, and the data is provided to external uses as .CSV 4 servers: 1 master database server, 1 live mirror, 1 web database server (updated hourly), 1 standby web server	1 - The data are not currently stored in a database system, but there are plans to move it to SQL server 2005. 2 servers: Main file server and file mirror at present	1 - The data are not currently stored in a database system, but there are plans to move it to SQL server 2005. 2 servers: Main file server and file mirror at present

Criteria	Ranking						
	NAEI	PCM	AURN	LAQN - Continuous AQ_DG	LAQN - MET	LAQN - Metals	LAQN - Traffic
INSPIRE	1 – The dataset does not adhere to INSPIRE standards and does not use XML	1 – The dataset does not adhere to INSPIRE standards and does not use XML	1 – The dataset does not adhere to INSPIRE standards and does not use XML	1 – The dataset does not adhere to INSPIRE standards	1 – The dataset does not adhere to INSPIRE standards	1 – The dataset does not adhere to INSPIRE standards	1 – The dataset does not adhere to INSPIRE standards

Table A6.2 Non-automatic Scoring Table:

Criteria	Ranking					
	NPL: Non-Automatic Hydrocarbon Network (HCN)	NPL: UK Heavy Metals Network	NPL: UK Black Carbon and Black Smoke Network	NPL: UK Particle Number and Speciation Network	CEH: Rural Heavy Metals Monitoring Network	CEH: EMEP (Auchencorth)
Data Purpose	Legislative network. Compliance with Benzene Air Quality Limit Value. 2008/50/EC and contains the following fields: Site, period of measurement, concentration, units	Legislative network. Compliance with Nickel, cadmium Arsenic and Lead Air Quality Limit Values. 2004/107/EC and contains the following fields: Site, period of measurement, concentration, units	Research network and contains the following fields: Site, period of measurement, concentration, units	Research network and contains the following fields: Site, period of measurement, concentration, units	<p>The data is collected for Defra's rural heavy metals monitoring network and is used for:</p> <p>EU Air Quality Framework Directive EMEP Protocol OSPAR Convention</p> <p>The data collected is concentrations of 26 separate heavy metals in air (PM10) fraction, and rainwater collected from 15 sites, temporal information</p>	<p>The data are collected for Defra's EMEP and is used for:</p> <p>the Co-operative Programme for Monitoring and Evaluation of the Long-range Transmission of Air Pollutants in Europe Policy</p>

Criteria	Ranking					
	NPL: Non-Automatic Hydrocarbon Network (HCN)	NPL: UK Heavy Metals Network	NPL: UK Black Carbon and Black Smoke Network	NPL: UK Particle Number and Speciation Network	CEH: Rural Heavy Metals Monitoring Network	CEH: EMEP (Auchencorth)
Data Searchability	1- The data are uploaded to the AQ archive. There is no search engine, the navigation is not immediately intuitive, however once the user gets accustomed to the site, it is very easy to navigate and find the desired dataset.	1- The data are uploaded to the AQ archive. There is no search engine, the navigation is not immediately intuitive, however once the user gets accustomed to the site, it is very easy to navigate and find the desired dataset.	1	1	2 – The data can be found through the use of an interactive map or	2 – Very simple online tool via the UK – pollutant deposition website, where the user only has to select the year and pollutant and they receive the suitable data.
Data Timeline	2 – The data are uploaded every 3 months with fortnightly measurements.	2 – The data are uploaded annually with weekly measurements.	2 – The data are updated annually with hourly measurements.	2 – The data are updated annually with hourly Nitrate, Particle number, particle size spectrum measurements and daily Organic and elemental carbon, sulphate nitrate and chloride measurements.	1 – The data are collected annually. The temporary resolutions of the data are weekly and annual mean.	1 – Data is updated annually but this is to be changed to quarterly. The temporary resolutions of the data are 20 minutes to weekly/monthly.

Criteria	Ranking					
	NPL: Non-Automatic Hydrocarbon Network (HCN)	NPL: UK Heavy Metals Network	NPL: UK Black Carbon and Black Smoke Network	NPL: UK Particle Number and Speciation Network	CEH: Rural Heavy Metals Monitoring Network	CEH: EMEP (Auchencorth)
Data Downloadability	2 – The user can download the data from the AQ archive.	2 – The user can download the data from the AQ archive.	2 – The user can download the data from the AQ archive.	2 – The user can download the data from the AQ archive.	0 – The data cannot be downloaded for heavy metals. The user can view the data through an interactive graph. This option is available for other datasets supplied via the website.	3 – The user can download the data from the site after entering the parameters they wish.
Data Historical Trending's	2 – 2001 to present	1 – 2004 - present	1 – 2006 - present	1 – 2005 - present	1 – 2004 – to present	1- 2006 - current
Data Comprehensiveness	2- The data are collected over national geographical coverage. The spatial resolution of the data is monitoring sites.	2- The data are collected over national geographical coverage. The spatial resolution of the data is monitoring sites.	2- The data are collected over national geographical coverage. The spatial resolution of the data is monitoring sites.	2- The data are collected over national geographical coverage. The spatial resolution of the data is monitoring sites.	2 – The data are collected over a national geographical coverage. The spatial resolution of the data is point source data collections extrapolated to national UK 5km maps	2- The data are collected over a national geographical coverage. The spatial resolution of the data is point source. Some of the datasets contain more readings, i.e. 6 with easting =0 and 3 where easting =0.

Criteria	Ranking					
	NPL: Non-Automatic Hydrocarbon Network (HCN)	NPL: UK Heavy Metals Network	NPL: UK Black Carbon and Black Smoke Network	NPL: UK Particle Number and Speciation Network	CEH: Rural Heavy Metals Monitoring Network	CEH: EMEP (Auchencorth)
Data Accuracy	2 - Site audits, data quality circle are performed on the data.	2 - Site audits, data quality circle are performed on the data.	2 - Site audits, data quality circle are performed on the data.	2 - Site audits, data quality circle are performed on the data.	3 - Calibrating of Instruments (every 6 months). Lab procedures are fully accredited (UKAS) with QA and inter-lab comparisons.	2 - Standard QA of data, graphing comparisons, + instrument calibrating
Data Consistency	5 - the data contains the same approach/ methodology across the whole time series	5 - the data contains the same approach/ methodology across the whole time series	5 - the data contains the same approach/ methodology across the whole time series	5 - the data contains the same approach/ methodology across the whole time series	5 - the data contains the same approach/ methodology across the whole time series (Unable to view – to confirm)	5 - the data contains the same approach/ methodology across the whole time series
Data Processing	2	2	2	2	2	2

	Ranking					
Criteria	NPL: Non-Automatic Hydrocarbon Network (HCN)	NPL: UK Heavy Metals Network	NPL: UK Black Carbon and Black Smoke Network	NPL: UK Particle Number and Speciation Network	CEH: Rural Heavy Metals Monitoring Network	CEH: EMEP (Auchencorth)
Metadata	3 - Sites collocated with AURN.	3 - Sites collocated with AURN.	3 - Sites collocated with AURN.	3 - Sites collocated with AURN.	2 –Site details, units, date range (Unable to view – to confirm)	2 - Contains the pollutant name, pollutant type, units, year and resolution.
Data Format and Standards	2 – The format of the dataset is .CSV.	2 – The format of the dataset is .CSV.	2 – The format of the dataset is .CSV.	2 – The format of the dataset is .CSV.	0 – The data is displayed in a graphical format; with no clear way no download.	2 - Internal: Excel/Logger Files. External: Pollutant Deposition website (Oracle)
INSPIRE	1 – The dataset does not adhere to INSPIRE standards and does not use XML	1- The dataset does not adhere to INSPIRE standards and does not use XML	1- The dataset does not adhere to INSPIRE standards and does not use XML	1- The dataset does not adhere to INSPIRE standards and does not use XML	1- The dataset does not adhere to INSPIRE standards and does not use XML	1- The dataset does not adhere to INSPIRE standards and does not use XML

Appendix 7

INSPIRE, SEIS and data.gov.uk

The INSPIRE Directive and SEIS principles are now developing from the largely conceptual and high level standards to more specific implementing rules. This detail will allow systems to be designed and implemented with datasets and metadata that comply with these agreed open and transparent standards, including those for handling spatial data, and will leverage the Semantic Web and use of Linked Data. These are explained in more detail by data.gov.uk¹³:

“The Semantic Web is an evolution of the World Wide Web that, rather than just linking from one document to another, focuses on their meaning in relation to each other. Linked Data is a set of technologies to achieve this for data, creating a web of data.”

INSPIRE Requirements and Annexes

INSPIRE requires that Member States must:

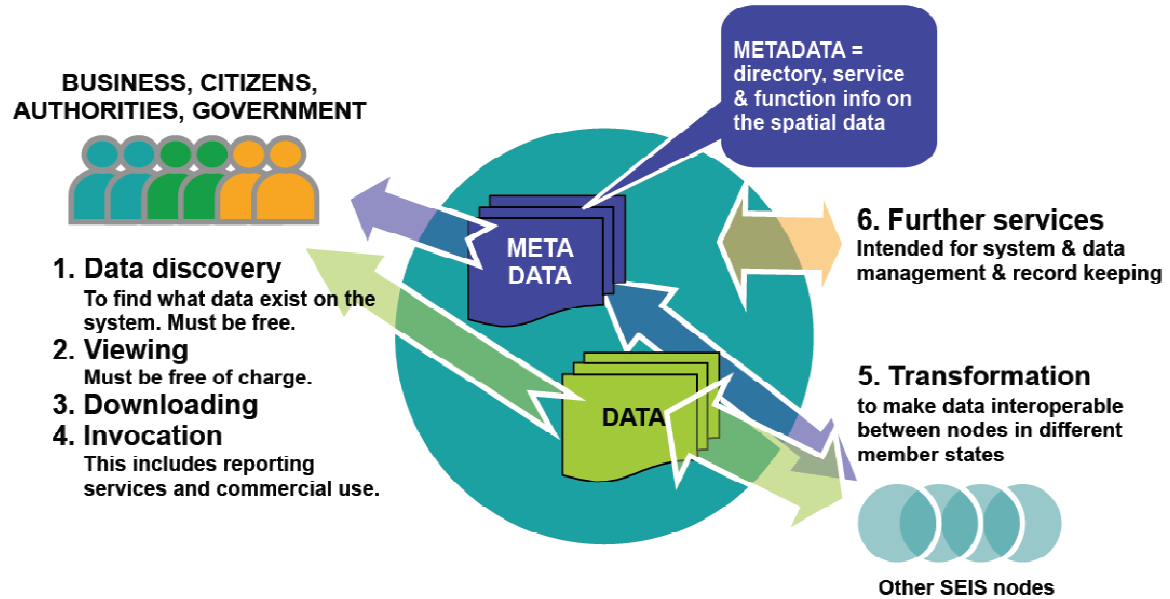
- Build infrastructures and network services for spatial information, made compatible by common Implementing Rules (IR) and Community Measures, that allow data exchange at EC and trans-boundary level
- The data must be stored at the most appropriate level, but data collected at one level of public authority must be available to all others
- Access rules must not unduly restrict data use, but charges for Rights Managed (RM) data is permitted, and Intellectual Property (IP) is protected
- The directive applies to all spatial data held electronically by or on behalf of public authorities and – subject to certain conditions – spatial data held by other natural or legal persons if they request it
- Practically all data that has a spatial component is affected (see next page)
- To speed data discovery, Member States must provide metadata on the data stored
- Member States must provide free public data discover and viewing services
- The Member States' infrastructure should be accessible through the EU INSPIRE geoportal

Figure A7.1 shows the link between INSPIRE and SEIS. The proposed integrated system for air quality data in the UK would be compliant with these.

¹³ <http://data.gov.uk/faq#whatisthesemanticweb>

Figure A7.1 INSPIRE and SEIS linkages

INSPIRE REQUIRES SIX FUNCTIONS OF SEIS NODES



The below table shows the different thematic areas covered by the different INSPIRE annexes. Those in bold are directly applicable to the high priority air quality datasets.

Table A7.1 INSPIRE Thematic Areas

Annex I	Annex II	Annex III
1 Coordinate reference systems 2 Geographical grid systems 3 Geographical names 4 Administrative units 5 Addresses 6 Cadastral parcels* 7 Transport networks 8 Hydrography 9 Protected sites	1 Elevation 2 Land cover 3 Orthoimagery 4 Geology	1 Statistical units 2 Buildings 3 Soil 4 Land use 5 Human health and safety 6 Utility and governmental services 7 Environmental monitoring Facilities 8 Production and industrial facilities 9 Agricultural and aquaculture facilities 10 Population distribution and demography 11 Area management/restriction/regulation zones & reporting units 12 Natural risk zones 13 Atmospheric conditions 14 Meteorological geographical features 15 Oceanographic geographical features 16 Sea regions 17 Bio-geographical regions 18 Habitats and biotopes 19 Species distribution 20 Energy Resources 21 Mineral Resources

* Cadastral parcel is a land registry entry, with attached data on ownership, user, rights and restrictions, localization, administrative boundaries, buildings or parts of buildings and all kinds of constructions, official zoning, land use, land cover, values/level of productivity, address(es) and description. In French and Italian the "cadastre" is the national land registry)

Ideally the proposed datasets to be used in the integration would all adhere to INSPIRE regulations, however it should be noted that the regulations for air quality data have not all been set, and that some of the air quality data falls under Annex II. Therefore the system should be adaptable to fit into the regulations from the new Annex II once it is released.

INSPIRE Implementation Timetable

Table A7.2 INSPIRE Timetable

Milestone date	Article	Description
15-May-2010	21§1 21§2	Implementation of provisions for monitoring and reporting
30-Nov-2010	15	The EC establishes and runs a geo-portal at Community level
03-Dec-2010	6(a)	Metadata available for spatial data corresponding to Annex I and II
19-Oct-2011	16	Discovery and view services operational
June 2012	16	Download services operational
June 2012	16	Transformation services operational
June 2012	7§3, 9(a)	Newly collected and extensively restructured Annex I spatial data sets available
03-Dec-2013	6(b)	Metadata available for spatial data corresponding to Annex III
January 2015	7§3, 9(b)	Newly collected and extensively restructured Annex II and III spatial data sets available
June 2017	7§3, 9(a)	Other Annex I spatial data sets available in accordance with IRs for Annex I
30-May-2019	7§3, 9(b)	Other Annex II and III spatial data sets available in accordance with IRs for Annex II and III

SEIS

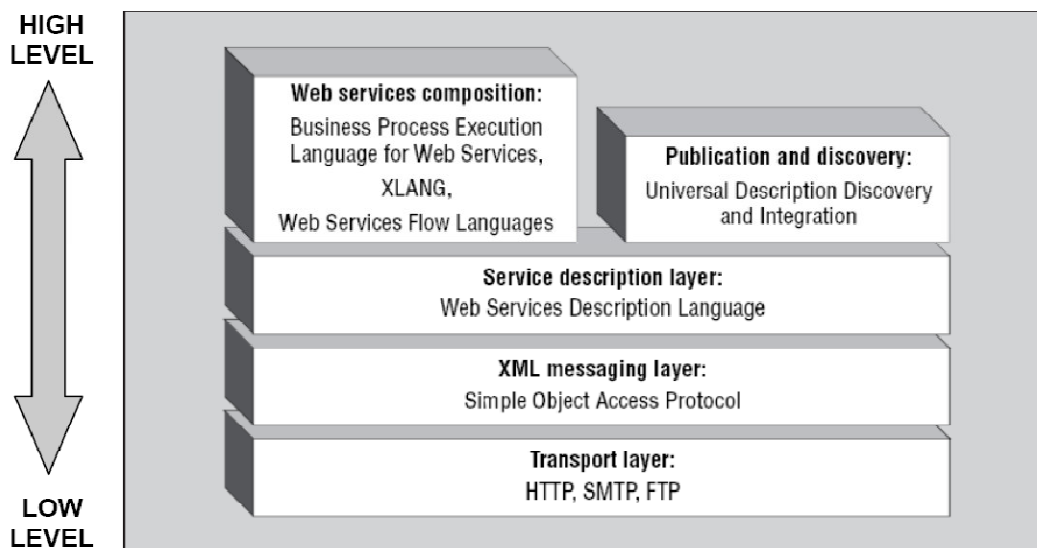
Much like INSPIRE (as INSPIRE is a component of SEIS) each Member State is required to implement network services that conform to the Implementing Rules (IR), which specify the general architecture, security, multilingualism, compliance modes, technical architectures and end user needs.

The SEIS is the resulting geospatial information sharing service, comprising a prescribed technical architecture and rules for:

- Data transport, based on HTTP, SMTP & FTP. The Hypertext Transfer Protocol (HTTP) is an Application Layer protocol for distributed, collaborative, hypermedia information systems. Simple Mail Transfer Protocol (SMTP) is an Internet standard for e-mail transmission across Internet Protocol networks. File Transfer Protocol (FTP) is a standard network protocol used to exchange and manipulate files over the Internet.
- Metadata & XML messaging, based on SOAP, Dublin Core. XML (Extensible Markup Language) is a set of rules for encoding documents electronically. SOAP (Simple Object Access Protocol) is a protocol specification for exchanging structured information in the implementation of Web Services in computer networks. The Dublin Core metadata element set is a standard in the fields of library and computer science. It is intended to be used for cross-domain information resource description. It defines conventions for describing things online in ways that make them easy to find.
- Services, based on Web Service Description Language (WSDL);
- Publication and discovery, using UDDI (online business yellow pages).
- Web service composition, with options of BPELWS, XLANG and WSFL. Business Process Execution Language for Web Services provides a means to formally specify business processes and interaction protocols. XLANG is an XML-based extension of Web Services Description Language (WSDL). WSFL (Web Services Flow Language) is an XML language to describe the composition of Web services.

The following diagram illustrates how these components sit in the technology stack.

Figure A7.2 SEIS Technology



Open Data Storage Formats

The future platform would make extensive use of XML and XML based technologies for data storage and exchange. Over time internationally standardised XML Schemas will be developed for the storage of air quality and other environmental data (largely through the INSPIRE Annexes). Alongside the standardisation for air quality data similar approaches will be applied to other datasets (at the very least due to the driver of data.gov.uk, possibly via INSPIRE depending on the nature of the data) allowing for non-air data (such as traffic data or health data) to be used far more easily alongside air data. This is particularly relevant to any spatial datasets as these will have to be compliant with INSPIRE in the future.

Metadata

As well as the data format the associated metadata (data that describes the nature of data) needs to be standardised. INSPIRE specifies the following for metadata and will include the following categories:

- Dataset title, Abstract, language, Dataset reference date, Dataset topic category
- Dataset responsible party
- Spatial resolution of the dataset, Reference system, Additional extent information for the dataset
- Data quality
- Distribution format and term and condition to use the data

The following standards are referenced at http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/metadata/MD_IR_and_ISO_20090218.pdf and need to be considered when implementing INSPIRE compliant metadata:

ISO 19115:
EN ISO 19115:2005, Geographic information - Metadata
ISO 19115/Cor.1:2006, Geographic information – Metadata, Technical Corrigendum

ISO 19119:
ISO 19119:2005, Geographic information - Services
ISO 19119:2005/Amd 1:2008, Extensions of the service metadata model

ISO 19108:
EN ISO 19108:2005, Geographic information – Temporal Schema

ISO 639-2
Codes for the representation of names of languages - Part 2: Alpha-3 coded control

ISO 8601
Data elements and interchange formats - Information interchange – Representation of dates and times

ISO/TS 19139
2007 Geographic information - Metadata – XML Schema Implementation

CSW2 AP ISO
OpenGIS Catalogue Services Specification 2.0.2 - ISO Metadata Application Profile, Version 1.0.0, OGC 07-045, 2007

ISO 10646-1
Information technology — Universal Multiple-Octet Coded Character Set (UCS)— Part 1: Architecture and Basic Multilingual Plane

Temporal Metadata

An important element to consider is the use of a standard to describe the temporal association with data (as well as the spatial which is the key focus for INSPIRE) so that data can be easily linked across different datasets for the same time period. The ISO 8601 standard should be used for representation of data and time. However limitations have been identified by Bordogna et al.¹⁴ which may need to be overcome depending on the nature of time data that are stored in a dataset.

Spatial Data

Data sets will need to include INSPIRE compliant spatial referencing systems and spatial encoding – for example using standardised coordinate systems¹⁵. This means that any data that are currently attributed to a geographic point will need to make use of a standard geographical referencing or encoding system. Whilst it will not be the mandatory encoding system; GML (ISO 19136) will be the default geographic encoding standard for INSPIRE. Further details are available in the Guidelines for Encoding of Spatial Data.¹⁶

Systems will also need to transform data that is currently stored against one coordinate system and make it available with other coordinate systems (for example if data are stored in Ordnance Survey grid references they also need to be available in Lat/Long). These are outlined by INSPIRE implementing rules¹⁷ on coordinate transformation services.

Semantic Web and Linked Data

Much like how web pages and documents (such as this one) reference their sources using references to URLs it is possible to do something very similar with datasets. The Linked Data Initiative <http://www4.wiwiw.fu-berlin.de/bizer/pub/LinkedDataTutorial/#intro> outlines how a standard called RDF (Resource Description Framework) is used to provide structure to information published on the web and to interlink data from different data sources. This approach allows end users to navigate between related data sources and for systems it allows for far better structure and standardisation and makes building models and managing data far easier.

¹⁴ Extending INSPIRE Metadata to imperfect temporal descriptions, <http://www.gsdi.org/gsdiconf/gsd11/papers/pdf/235.pdf>

¹⁵ D2.8.1.1 INSPIRE Specification on Coordinate Reference Systems – Guidelines, September 2009
http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_Specification_CRS_v3.0.pdf

¹⁶ http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/DataSpecifications/D2.7_v3.0.pdf

¹⁷ Draft Technical Guidance for INSPIRE Coordinate Transformation Services, October 2008
http://inspire.jrc.ec.europa.eu/reports/ImplementingRules/network/Draft_Technical_Guidance_Coordinate_Transformation_Services_v1.0.pdf



Didcot
Oxfordshire
OX11 0QJ

Tel: 0870 190 6602
Fax: 0870 190 6377