Innovation Development using a Semantic Web Platform: A case-study from the National Health Service in England

Mark Birbeck[1], Dr Michael Wilkinson[2]

[1]Sidewinder, 2nd Floor Titchfield House, 69-85 Tabernacle Street, London, EC2A 4RR, UK. mark.birbeck@webbackplane.com
[2]NHS National Innovation Centre, New Kings Beam House, London SE1 9BW michael.wilkinson@institute.nhs.uk

Abstract: The National Innovation Centre, which is part of England's National Health Service, has developed a Semantic Web Platform to improve the speed and quality of health technology innovation development. This platform links data from multiple sources, via automated, semi-automated and manual processes. The collected data and information is rendered on the NIC's Triple Store using the NIC's Open Health Innovation Ontology (OHIO). Powered by Solr, OHIO-based results generate highly context specific information for end-users. End-users are able to visualise data via a range of open-source visualisation tools, and these visualisations can then be exported and re-deployed on new platforms via the NIC's dataShuttle widget.

Keywords: health, innovation, linked data, OHIO, dataShuttle, RDFa, NHS National Innovation Centre

1 About the National Health Service

The National Health Service (NHS) in England was established in 1948, and is the World's largest publically funded health service. The NHS provides comprehensive, universal, and free at point of delivery healthcare for the entire population of England, representing some 51 million people. On average, the NHS serves one million people every 36 hours.

The NHS is also world's 4th largest employer, and employs 1.7 million people, including:

- 120,000 hospital doctors
- 40,000 general practitioners
- 400,000 nurses
- 25,000 ambulance staff
- 55,000 scientists

2 UK Research Environment

The UK Life Sciences industry is a world-leading, high-tech industry employing over 120,000 people and investing at least £4.6 billion in research and development in the UK. It is a strong driver of economic growth, provides highly-skilled employment opportunities and, through the development of innovative medicines and medical technologies, contributes to the delivery of high-quality healthcare.

In addition to industry, the UK university sector is particularly strong, and includes the University of Cambridge, University College London, Imperial College London, and University of Oxford.

3 The NHS National Innovation Centre

The NHS National Innovation Centre (NIC) was established by the Department of Health in September 2006 in response to Industry's complaint to Government that the NHS was unresponsive to innovation. In the current economic climate, Government, NHS and Industry all see innovation as a key driver to improve NHS productivity and quality of care. Innovation is also seen to support wealth and job creation, inward investment and export trade

The NIC seeks to achieve its aims to speed-up the development and delivery of technological innovations likely to benefit the NHS and UK industry by awarding research and development grants and contracts. To support its work, the NIC has developed a Web 2 innovation management infrastructure that is ISO9001 certified. This web-based infrastructure is based on a staged innovation work-flow that starts with defined need, and then leads to design, development, demonstration and diffusion of an innovation into the market.

Although the NIC's Web 2 infrastructure has helped to improve operational efficiency, it has proven sub-optimal. To explain, innovators and those who support innovation need access to high quality, context specific data and information in order to make informed decisions. Search tools such as Google offered very little utility to innovators and those who support them. In short, there was a pressing need to improve access to high-quality and useful information in order to improve efficiency, effectiveness and quality of care in the NHS in general, and innovation development in particular.

Recognising the potential of Linked Data, in January 2010 the UK Government launched data.gov.uk in order to make public data public. Data is increasingly being made available via data.gov.uk; for example, the Department of Health has contributed over 2,760 datasets, and increasingly these datasets are in formats that are machine readable, such as CSV files. For the NIC, data.gov.uk is an important source of data that has the potential to serve as a catalyst for innovation development. However, whilst it's important to publish information, it's also crucial to be able to make use of it: Linked Data is necessary but not sufficient.

The challenge that the NIC has faced is essentially to create a *semantic* capability that exploits Linked Data. To achieve this, the NIC continues to develop a range of widgets and applications for use in the field of healthcare technology that allow data to be viewed and manipulated in a variety of ways. In order to put these tools on a solid foundation, the NIC has developed a modular, extensible OWL ontology that describes the process of innovation and those involved in it.

This OHIO ontology, which is described in more detail below, has been designed in such a way that it will also provide the basis for describing innovations more broadly, not just those involving healthcare technologies.

The NIC is committed to using Semantic Web technologies as a way to improve the speed and quality of decision making in the area of health technology innovations. The semantic web enhancements of the NIC toolsets will be used to assist:

- decision-making clinicians, managers and commissioners in the NHS;
- healthcare technology industries;
- other government departments

At the same time, the NIC is committed to contributing back to the community from which these technologies have come, and is therefore making its platform available in the form of the Semantic Web Platform.

4 THE NIC's Semantic Web Platform

The key components of the NIC's Semantic Web Platform are discussed below.

4.1 OHIO: Open Health Innovation Ontology

Increasingly over the past decade, syntactic search tools such as Google delivered results of little utility to both innovators and those who support innovators. To put it simply: very large search results could be generated very quickly, but the results are largely of little relevance.

In order to enable the NIC to meet its aim of speeding-up the development and diffusion of health innovations likely to benefit the NHS and industry Provide innovators information that has:

- **Context:** health technology innovation
- **Coherence:** Information makes logical sense, within context

- **Quality:** valid and reliable information (i.e. provenance, curation)
- **Freshness:** up-to-date information, available whenever and wherever
- **Purpose:** found information can be shared and exported, contributing to the advancement of knowledge and development of assets

Following on Tim Berners-Lee's presentation on Linked Data at TED (March 2009), the NIC began work on creating a semantic platform to support healthcare technology innovation development. The NIC began creating ontology for innovation development. The aim of this work was to identify and define the essential entities of health technology innovation and the relationships between such entities, as grounded in the experience of elite innovators. Having achieved this, the NIC would then convert the ontology according to W3 standards so that the ontology could be used by the NIC, its stakeholders, and others on the internet.

The research project was managed by Dr Michael Wilkinson of the NIC. Dr Dirk Wierich designed and conducted the study, which was based on Grounded Theory methodology. A sample of 32 individuals participated in a series of semi-structured and open-ended interviews during the summer and autumn of 2009. This sample comprised of experienced innovators and clinicians who were judged by their peers to be at the vanguard of healthcare innovation.

The interview results revealed clustering of thoughts around a number of clearly defined entities, as well as relationships between these entities. A written structure of the innovation ontology was created by Dr Wierich. This written structure was translated and encoded by Mark Birbeck, using OWL and SKOS. The development of OHIO seeks to embrace and add-value to other ontologies where appropriate: OHIO is easily extendable; for example, OHIO incorporates FOAF, Dublin Core, DOAP and will soon include SWAN and Good Relations.

OHIO was then put on the NIC's Triple Store, where it received further testing before a successful release on the Internet. For a technical discussion and documentation about OHIO (Open Health Innovation Ontology), see: http://code.google.com/p/argot-hub/wiki/GettingStartedOhio

The faceted browser on the NIC's Semantic web platform is based on OHIO. Each facet is an OHIO entity, and the connections between the facets are the semantic expression of OHIO. Similar, the NIC's dataShuttle widgets are based on OHIO. These two NIC creations – OHIO and dataShuttle - should go some way to start meeting the need to enable clinicians and innovators quickly to search, find, export, and use data in a more meaningful way than having to find information via key-word searches via disconnected portals.

OHIO is essentially domain ontology. Given the robustness of OHIO, the NIC has also published Open Innovation Ontology (OIO) as core ontology. Specialist domains interested in innovation are encouraged to build on and use both OHIO.

4.2 Data Sources

There are an ever increasing number of publicly available data sources available on the internet. The NIC is harnessing the potential of multiple data sources to support its innovation platform.

For example, a major initiative in the UK is data.gov.uk. This freely available web resource makes public data public in order to spur innovation, and to serve as a catalyst to improve public services and to generate wealth through new forms of enterprise. The NIC uses both health and non-health datasets to support its work (for example, health data linked to school or environment data). The provenance of these datasets is of critical importance to innovation development. This provenance for health data rests with the NHS Information Centre, which is working collaboratively with the NIC.

Another important data source for the NIC is PubMed. PubMed comprises more than 20 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

Data can also be sourced from other websites via RDFa, which was invented by one of the authors, Mark Birbeck. RDFa is capability is essential, as it provides the opportunity for other websites to be enhanced by and contribute to NIC's Semantic Platform.

The datasets issued by data.gov.uk and PubMed are accessed and then rendered against the OHIO ontology via automated processes. To complement this automated process, the NIC also has a specialist team that conducts global horizon scanning of published information about health innovations. In this case, the information found is curated against the OHIO ontology and entered into the Triple Store using semi-automated processes developed by the NIC.

4.3 A Widget Platform

To help the NIC achieve its key goals of flexibility in use and agility in build, it has eschewed the traditional notion of an application. Instead it is building a platform that makes it possible to create different types of visualisation against different sets of data, and then to combine those visualisations onto a page. The platform is essentially a widget platform, allowing users to browse, categorise, rate, and discuss widgets and datasets, and then to deploy visualisations. By providing social networking facilities, it is easier for people to discover and share techniques and interesting insights on the data. It's also possible for the more technically minded to create different widget types and add them to the catalogue.

Once a data/widget combination has been chosen and configured (i.e. that is, once a visualisation has been created), the platform also makes it possible to deploy the live visualisation to the web, perhaps in a blog or journal article, or to embed a snapshot of the visualisation in a business case or report.

4.4 Applications as Visualisation Collections

Since the data in the platform can be presented in many different ways and used by many types of user, it's almost impossible to create a single application that will appeal to everyone. So instead the focus whilst building the platform is on creating as many visualisations as possible, and then allowing those visualisations to be combined in a variety of ways.

To make it easy to combine visualisations that are based on a common piece of data, the platform makes use of the W3C's XForms standard. XForms is more commonly used as a means of marking up data entry forms, but it also provides a number of key features for use in the platform. The first is a UI framework into which it's possible to drop custom visualisations; in essence each visualisation becomes an XForms control, and the XForms engine then becomes responsible for the control's lifecycle.

The second benefit that XForms provides is the dependency engine, which automatically keeps the state of all related controls up-to-date based on changes in other controls. This makes it easy to link a number of visualisations together based on a common value, such as a year, a type of treatment, or a particular country.

4.5 Architecture

Triple Store
The foundation of the entire architecture is a Triple Store. The core data about innovations, people and organisations is obviously held in this store, but so too is other information that has been imported from other sources, such as those on data.gov.uk. The store also holds categorisation information about widgets and data sources, so as to be able to help users create visualisations by matching data with appropriate widgets.

The Triple Store used could be one of a number of available Open Source stores, but for our work we've focused on using Virtuoso. There are a number of reasons for this, but the key benefits were that it's fast, it scales, and its feature set enables the importing of triples and querying them with SPARQL.

Natural Language Processing (NLP)
Sitting alongside the Triple Store is the NLP module, whose job it is to import information from other sources, such as medical journals, patient forums, news articles, and so on. Other organisations' websites are a key source of information for use in the platform. For example, there are over 50,000 scientists working in the NHS, and the only realistic way to keep the system up-to-date on the projects and research those people are undertaking, is to import data from their individual websites and those of their departments. The NIC is involved in a number of projects aimed to adding RDFa to key sites that can then provide useful information to the platform.

SPARQL end-point
The key interface to the Triple Store is the SPARQL end-point. This allows queries to be run over the web, against the data in the store. The SPARQL end-point will shortly be made available to other developers, for use in their applications, meaning that other departments and organisations will have access to the same data that is driving the NIC's own applications.

4.6 User Interface

User Interface management
The User Interface management layer of the platform is provided by Drupal 7. Using a Content Management System (CMS) seemed obvious, since there is no point in reinventing many of the components that are needed, such as forums, comments, ratings, user registration, and so on. However, Drupal was chosen specifically because of the well-known commitment to the Semantic Web of leading members of its development community.

Drupal's own data is stored in a traditional SQL database, but there are a growing number of modules that allow RDF and SPARQL to become part of the mix.

In terms of the functionality needed for the platform, each dataset and widget has its own page, which allows it to be a focus for comments, ratings and categorisation. Further pages are then available when users create visualisations via a combination of data and widget.

The OHIO Faceted Browser
The User Interface is based on a faceted browser. The facets are, essentially, the key elements of the OHIO ontology. As the end-user selects facets of interest, the listed results become more contextual and relevant to the needs of the end-user.

The faceted browser is supported by Solr. Solr is a fast, open source enterprise search platform from the Apache Lucene project. Its major features include powerful full-text search, hit highlighting, faceted search, dynamic clustering, database integration, and rich document (e.g., Word, PDF) handling. The NIC has found Solr to be highly scalable, providing distributed search and index replication. An additional benefit of Solr is that it powers the search and navigation features of many of the world's largest internet sites.

dataShuttle Visualizations
There are now an enormous number of high quality graphing and charting libraries available, each with their different strengths. Since this landscape is constantly changing and improving, it was important for the platform to be able to leverage the best of these components, as well as to be able to take advantage of new components as they become available. For this reason effort was put into ensuring that data from the SPARQL end-point could be consumed by components from libraries such as Google's Visualization API3 and MIT's Exhibit4.

However, even this is not enough, since the aim of the platform is to make it as easy as possible for authors to consume and display the data, and to build applications. The next step was therefore to wrap the widget and data combination into an XForms custom control. These controls have been built on top of backplanejs5, an Open Source JavaScript library which provides support for XForms and RDFa. To embed a Google Visualization map now requires nothing more than authoring a simple control:

```
<xf:output bind="mapdata" mediatype="visualization/GeoMap" />
```

The XForms foundation is important for another reason which is that its spreadsheet-like dependency engine allows a number of visualisations to be connected to the same data. For example, three visualisations might be showing different views on data for the year 2021, such as a map, a bar chart and a pie chart. The XForms dependency engine ensures that if the data is changed to 2021, all three visualisations will be notified and updated without the programmer having to do anything.

By combining visualisations on the same page, it is possible to use these components to create applications.

dataShuttle Gallery: Many NIC end-users are expert innovators, clinicians and scientists. When an end-user uses the OHIO facetted browser, they are usually looking for information to solve a particular question or problem. Through active searching and refinement, the end-user is able to find information, create visualisations and generate dataShuttles. These dataShuttles are important, semantically enriched assets that may be of use to others who have similar information needs. The dataShuttle Gallery is essential a place for innovators to profile their dataShuttles, and for others to rate and provide comments on dataShuttles.

5 The Future

The architecture described here has applicability beyond health, and even beyond innovation, and many organisations would benefit from having a ready-made platform such as this one. For this reason the NIC has initiated an Open Source project to create a set of guidelines, best practice and tools that it hopes will help anyone who wants to make their data part of the Linked Data cloud.

Everyone should benefit as more people join the project and contribute to the code and guidelines, and more organisations put their data into the Linked Data cloud.

The NIC's Semantic Web initiative provides a genuine opportunity to provide tools that will assist decision-makers and healthcare professionals, and to enhance the process of developing and capitalising on innovations. But building these tools on Open Source platforms and in a collaborative manner means that others can benefit from the work, as well as contribute to it.

As the work progresses the NIC hope to see a community grow around the platform, sharing tools and visualisations that make it easy to build applications. The NIC also hope to see an increasing number of user groups benefit from this work.

Acknowledgement: The authors wish to thank Brian Winn, Marie Maher, Jonathan Wong and Dr Nigel Sansom of the NHS National Innovation Centre for contributing to the drafting of this manuscript.

Appendix

Criteria	Evidence
Mandatory Features	
The application has to be an end-user application, i.e. an application that provides a practical value to general Web users or, if this is not the case, at least to domain experts.	The NIC's semantic web platform is designed specifically to support end-users, including innovators, clinicians and procurement officials. See Section 4, for full details. The NIC's semantic web platform is freely available on the Internet, via: http://datashuttle.nic.nhs.uk/
The information sources used • should be under diverse ownership or control • should be heterogeneous (syntactically, structurally, and semantically), and • should contain substantial quantities of real world data (i.e. not toy examples).	Information sources used in the application are under diverse ownership, are heterogeneous, and contain substantial quantities of real world data (e.g. data.gov.uk; PubMed). See Section 4.2
The meaning of data has to play a central role. • Meaning must be represented using Semantic Web technologies. • Data must be manipulated/processed in interesting ways to derive useful information and • this semantic information processing has to play a central role in achieving things that alternative technologies cannot do as well, or at all;	The meaning of data plays an essential role in the NIC's semantic web platform. Through a rigorous scientific process, the NIC created the Open Health Innovation Ontology (OHIO). This ontology represents the essential elements of innovation development and the connections between these elements. OHIO is the basis of the faceted browser, and ensures that data is presented to end-users is listed is highly contextual and meaningful. Using the OHIO faceted browser, end-users are able to create, share, and export highly contextual and meaningful visualisations. See Section 4
Additional Features	
The application provides an attractive and functional Web interface (for human users)	The NIC's semantic web platform has been designed by experts who specialise in creating easy-to-use web interfaces for platforms that are rich in data. See http://datashuttle.nic.nhs.uk/
The application should be scalable (in terms of the amount of data used and in terms of distributed components working together). Ideally, the application should use all data that is currently published on the Semantic Web.	The application is deployed on Virtuoso Triple Store and functions in the Cloud. This enables an efficient use of resources and full scalability. See Section 4.5 The application uses data that is currently available on the semantic web. See Section 4.2.
Rigorous evaluations have taken place that demonstrate the benefits of semantic technologies, or validate the results obtained.	The development of the Open Health Innovation Ontology (OHIO) was the result of extensive scientific study. OHIO, in turn, drives the Triple Store, faceted browser, and dataShuttles. See Sections 4.1 and 4.6

Novelty, in applying semantic technology to a domain or task that have not been considered before	The domain that the NIC operates in is health technology innovation. The NIC's Semantic Web Platform is designed to speed innovation development through the exploitation of semantic technology, including the creation of OHIO, and the use of open source technologies such as Drupal, Solr, open source visualisation tools, and the use of public data. See Section 4.
Functionality is different from or goes beyond pure information retrieval	The facetted browser is unique in that it is driven by OHIO. The dataShuttle is a unique widget that was created by the NIC to enable end-users to create, share and export data visualisations. See Section 4.6
The application has clear commercial potential and/or large existing user base	There are currently over 8,000 registered users of the NIC's innovation tools. The commercial potential exists as a result of improved performance in innovation development. Speed-to-market is essential in a competitive environment, and the Semantic Platform will go some way to improve performance. See Section 3
Contextual information is used for ratings or rankings	Solr is used to rank search result listings based on OHIO classifications. Solr enables the NIC to calibrate the results listings to provide more meaningful results to end-users. In addition, dataShuttles created by end-users are shared in a dataShuttle Gallery, where the provenance of the dataShuttles is recorded and where other users can add comments and rate the utility of each dataShuttle. See Section 4.6
Multimedia documents are used in some way	The Semantic Web Platform is designed to enable multimedia documents to be included as part of the end-user experience. These include text and data files, in addition to moving and static images. See http://datashuttle.nic.nhs.uk/
There is a use of dynamic data (e.g. workflows), perhaps in combination with static information	The NIC has developed a Web2 innovation management infrastructure that is ISO9001 certified that is based on a staged innovation work-flow that starts with defined need, and then leads to design, development, demonstration and diffusion. This platform has been enhanced to benefit from the potential of the semantic web. See Section 3
The results should be as accurate as possible (e.g. use a ranking of results according to context)	The OHIO ontology is the basis of the faceted browser and Solr search architecture. This combination of open source tools generates a highly context specific ranking of results for end-users. See Section 4.6.