**NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE**

**Linked Data project: Initial findings**

# Abstract / Summary

Linked Data technologies and methods for modelling and managing content have the potential to address many of NICE’s current challenges in the development, maintenance and dissemination of guidance. They are now approaching a level of maturity and penetration that makes them difficult to ignore. However there remains an element of risk when considering the adoption of the more novel aspects of linked data. This project aims to mitigate that risk through empirical study within the context of NICE’s information challenges.

A basic ontological model of NICE content has been created and populated with a subset of NICE content. The model and content have been used along with a rudimentary set of Linked Data technologies to create a demonstrator to test a number of hypotheses and real-world scenarios. This work and the resulting demonstrator has allowed the project team to learn and to convey some of the potential benefits in a tangible way to all interested parties across the organisation, in particular to those working to define a NICE content strategy. An initial set of observations have been made by the team working closely with both the data model and technology. Broadly the project team feels that while there is a steep learning curve associated with elements of a linked data solution, it is achievable and beneficial to work with NICE content in this more formally structured way.

# Purpose

This report presents the initial findings from the Linked Data project to indicate the progress made to date and to share early observations. The demonstrator developed in the course of the project will be presented in September 2014 and detailed findings will be delivered by early October 2014.

# Background

This project is a research initiative that falls under the aegis of the Guidance Development Project. It is intended to inform elements of the NICE content strategy, and the Blueprint which aims to set out the transition to structured authoring of NICE guidelines. The project began in April 2014 and will run until 30 September 2014

# Methods/Approach

Five hypotheses were devised to test the ability of linked data to address the kinds of problems with which NICE is currently faced. Under each hypothesis at least one real world scenario capturing users’ information needs was developed and used to guide the project. The scenarios were subsequently validated where possible with subject matter experts to elicit actual information needs at each point and a series of queries were documented that the persona in the scenario would require in an ideal world to address the particular scenario.

A demonstrator was developed to support several core needs outlined by the scenarios. The demonstrator has involved work in each of the following areas:

* 1. Modelling
     1. Working with experts across NICE to understand how NICE guidelines are constructed
     2. Understanding and validating the individual core components of guidelines and quality standards and the meaningful relationships between those components.
     3. . This model has been more formally expressed using Web Ontology Language (OWL)[[1]](#footnote-1) to create a core NICE Ontology.
     4. Incorporating external ontologies such as PROV[[2]](#footnote-2) and Open Annotation[[3]](#footnote-3) to model provenance and annotations.
  2. Data/Content
     1. Manual extraction of a subset of content from existing guidance and quality standards focusing on the topic of Diabetes
  3. Technology and User interface
     1. Use of ontology modelling tools (Protégé)
     2. Development of core technologies to build models
     3. Configuration of a content (triple) store to hold both the model and data,
     4. Automation of the loading of content into the triple store
     5. Automation of concept extraction from core NICE content entities using a natural language processing (NLP) tool
     6. Linkage of concepts in the content to existing external datasets and vocabularies such as Freebase[[4]](#footnote-4), MeSH[[5]](#footnote-5) and DrugBank[[6]](#footnote-6).
     7. Development of standard SPARQL queries to retrieve data to support scenarios
     8. Creation of a user friendly presentation layer
     9. Enable access to the data in graph form through the presentation layer to allow users both to:
     10. Exploration of a simple forms based approach for creating and uploading content into the graph database
  4. Communications
     1. Creation of a [blog](https://nicelinkeddata.wordpress.com/)[[7]](#footnote-7) to discuss the background to the project and specific issues as they arose

# Discussion of initial findings and observations

Detailed observations captured on the various stages of work leading to creation of the demonstrator are captured in Appendix 1. Core observations and observations pertinent to specific hypotheses are listed below, with links to detailed observations in the appendix where relevant.

* 1. **Core observations** 
     1. Modelling ontologies and developing the technology is possible and proven
     2. The process of modelling is complex, from choosing the type of modelling language, formality, logic and reasoning to defining the line between content, data and metadata.
     3. The choice of annotating blocks of content versus modelling fundamental concepts and relationships needs to be taken case by case, weighing up maintenance and tagging effort versus the benefits of a more granular model.
     4. Provenance and FOAF[[8]](#footnote-8) ontology models are some of the most robust existing ontology standards and should be incorporated into the NICE model where appropriate.
     5. Getting the model right is fundamental to the success of implementing linked data principles and will require the right level of resource.
     6. NICE Content cannot be simply separated and loaded into a formal model as it is. A level of re-authoring will be required.
     7. The level of structure required to make guidance ‘executable’ i.e. to enable fundamental integration within decision support tools, is significantly more than just to support authoring and presentation enhancements.
  2. **Observations on Hypothesis 1:** *NICE can efficiently and effectively develop content that is structured and embedded with rich metadata*
     1. A rudimentary method for writing/creating content into the graph database has successfully been created as part of the demonstrator, illustrating the potential to author content against specific ontologies and schemas.
     2. The Resource Description Framework (RDF) and its serialisation alone does not deliver what we might consider as ‘fully structured’ content. It is common to use some form of markdown or XML schema to impose lower level structure on the content i.e. for tables, bullets/steps etc.
     3. To the project teams knowledge no robust authoring tools exist that support both text editing and rdf structure, with the implication that the development or non-trivial configuration of an authoring tool would be required.
     4. Assigning terms and concepts to content during the authoring process is likely to be more efficient than retrospective tagging. However this was achieved retrospectively through use of the NLP tool Textrazor[[9]](#footnote-9) (observation 8.2) This tool or something similar could be used to identify concepts in text at the time of authoring. The demonstrator illustrates how this might work.
     5. The automatic identification of terms within a block of text, and the associations within existing published data sets shouldn’t be relied upon to be 100% accurate. However they appear accurate enough to drive suggestions that could then be verified by a subject matter expert or author.
     6. The ‘NICE ontology’ has modelled a number of relationships that are relevant to a recommendation and in a full solution these would be suggested to the author at the time of creation of content to enable linkage of relevant content (e.g. evidence statements, discussion, studies). The demonstrator does not show this, but by the end of the project, a mock-up will be developed to show how this might look.
     7. The demonstrator shows that it is possible to link to concepts in content to existing external datasets/vocabularies such as Freebase, MeSH and Drugbank. NICE will need to consult on which are the most appropriate external vocabularies to link to. (Observation 8.3)
     8. If internal NICE terms taken from the existing NICE vocabulary pool, or something similar, are used to annotate content, these will themselves need to be modelled using the SKOS ontology for them to have formally defined meaning outside of NICE and support a move towards semantic interoperability of our content (Observation 8.4)
  3. **Observations on Hypothesis 2:** *NICE can maintain information products that are embedded with rich metadata efficiently and effectively and understand the impact of changes in the evidence base on the NICE portfolio of information products, including the potential significance of impacts on published guidance.*
     1. The NICE ontology has successfully modelled relationships between quality standards/statements, guidelines, recommendations, evidence statements and studies and the dataset can be queried to show the impact of a particular study/studies on associated NICE content.
     2. The annotation of content at the evidence statement/recommendation level with concepts taken from standard vocabularies enables the data to be queried to see the recommendations/evidence statements that are about a particular topic area (such as treatment of diabetes with a particular combination of drugs). If the evidence in this area were to change, the affected evidence statements, recommendations and other products that are supported by those recommendations should be easily identifiable and thus the potential impact easily assessed.
     3. Understanding the wider changes in the evidence base is a large and complex problem involving the systematic review process. Since Cochrane are currently working on addressing this issue through the detailed modelling of studies and evidence using linked data, the NICE ontology only models studies in a basic way for the purposes of this project. NICE can benefit from the work that Cochrane has already undertaken in this area and collaborate with Cochrane in the future to ensure the work is of mutual benefit. It is not recommended for NICE to tackle this complex area in isolation. Opportunities to collaborate more closely with Cochrane have been identified.
     4. Updating of guidance iteratively to reflect changes in the evidence base will require sophisticated temporal modelling which has not yet been implemented in the ontology (observation 4.7)
  4. **Observations on Hypothesis 3:** *NICE will be better able to provide its content to our intended audience in a form that will enable users to access the right information at the right time, using content that is embedded with rich metadata including:* 
     1. The demonstrator shows that through annotation of content at the level of recommendations/evidence statements, a user will be able to find the specific content that they are seeking
     2. Breaking out the guideline into its constituent parts and assigning a recommendation a unique identifier as has been a part of this project, enables a source such as CKS to be able to link to NICE recommendations at a more granular level rather than at the guideline level which currently increases the time required to get to the specific information a user requires.
     3. The demonstrator shows the possibility of linking concepts (e.g. Diabetes) in the recommendations to concepts in other sources. For this project we have linked to DrugBank, as an example of how a query could be formed to move from recommended drugs to see their contraindications. Drugbank does not hold the same data about drugs as BNF, but is used to demonstrate how BNF data could be queried if the data were structured in the appropriate way.
  5. **Observations on Hypothesis 4***: NICE can better integrate its content with other information systems and third party systems when embedded with rich metadata enabling linking of data.*
     1. By increasing the level of formality used to model NICE content and associated metadata it is clear that other web applications and search engines can more easily identify and ‘understand’ NICE content. Conversations have taken place with NHS Choices (who are running a similar linked data initiative) to understand the implications of both parties publishing and integrating their content using linked data.
     2. To understand how NICE needs to provide its content for integration into third party systems such as decision support systems is more complex and will require continued working with third parties. The team has held several information sharing meetings with Manchester University to explain the aims of our project and there may be an opportunity to work with them as part of a project to explore this. Manchester have broad experience in working with companies such as Siemens to develop clinical decision support systems. Another opportunity is the “Encore” project led by the Universities of Birmingham and Oxford, currently seeking funding from the Health Foundation. This builds on the OpenClinical project which worked on development tools and techniques for building healthcare applications.
  6. **Observations on Hypothesis 5:** *The technologies available for the creation and maintenance of NICE information products embedded with rich metadata and for consumption of these products by the broader NHS system are mature enough for the adoption by NICE.*
     1. The demonstrator has been developed in a short space of time with a small team, implying that the technologies should be considered as mature enough for small scale / controlled adoption as part of a live service.
     2. Cost of ownership, implementation, security, fit with the technology stack and architectures have not yet been assessed and form part of the additional analysis which will need to be performed by the end of the project.
     3. NHS wide system adoption considerations will be achieved through further discussions with groups such as NHS Choices and HSCIC

# Early Conclusions (to be further validated prior to detailed findings)

* 1. Linked data concepts/technologies are mature enough to consider adoption, but will not be without challenges. However, it is believed that these challenges are just ‘different’ rather than necessarily more complex than those presented by alternatives (i.e. relational modelling, doing nothing etc).
  2. Whilst linked data is not a panacea and will not be appropriate universally, in order to meet the challenges NICE faces, the need to elevate the formality of the way that metadata and controlled vocabularies are defined and managed is clear and the need in certain circumstances to consider linked data principles (graph modelling) as a primary option over noSQL / relational data models for instance is also clear. It is believe that these principles should be documented and adopted by both the design authority (DA) and information standards working group (ISWG) to promote the appropriate use of this technology within NICE.
  3. We are looking for solutions to our challenges at a time when the sands are shifting quickly. While there may be benefit in waiting for more solutions to be developed, this needs to be weighed against the risk of inaction and the lack of NICE influence in solutions developed by others.
  4. The highest value least risk area to consider for initial adoption using Linked Data would be ‘people and groups’. While this will not deliver a complete solution to authoring all content in a structured way it offers the following advantages:
     + A well- established ontology (FOAF) to model people already exists and can be easily adopted by NICE
     + The problem space is well-defined
     + The management of information on people and groups currently in NICE, while not NICE’s core business, does present challenges.
     + Information about people involved in guidance development is an integral part of the content itself and could highlight the use of linked data for management of core content and support a reframing of guideline content as packages of discreet information that can be pulled together in different more efficient ways.
     + There are a number of projects in flight that could benefit from this modelling work (security model, in development, CRM / User rationalisation)
  5. We believe that there is significant benefit to continuing collaborative research in the areas of evidence management and interoperability for decision support to ensure NICE can both avoid duplication of effort and exert influence to support its own business goal.
     1. Potential collaborators in third party decision support / tool development are the OpenClinical Project, University of Manchester and NHS Choices
     2. Potential collaborators in the areas of evidence base monitoring and systematic reviews are: Cochrane, DECIDE/MAGIC programmes and the EPPI Centre

Next steps?

**Sponsor:**

Alexia Tonnel, Sarah Cumbers

**Authors:**

Andy Mitchell, Information Architect

Alison Desimone, Information Architect

**August 2014**

**Appendix 1: Detailed Observations**

| Area of focus | Observation | Implication |
| --- | --- | --- |
| 1. Observations on Content Extraction from Guidelines and Technology Appraisals | **1.1** Variation in structure and labelling of content in guidelines  The framework for extracting the concepts was based on replicating the structure of one guideline, initially CG15, incorporating the core elements common to all guidelines. However, the exercise was extremely time-consuming, as the other guidelines did not all fall neatly into the same pattern of labelling, section order or level of granularity as CG15, although the basic concepts of recommendation, evidence statements discussion were all present. There was a particularly wide variation in the labelling of “discussion” (evidence to recommendations discussion). The reasons for this are clear to understand: historical differences between types of guidance such as public health and clinical guidance, changes over the years in the template used and guidance produced by different groups in a different style. | The work currently being undertaken in NICE to standardise structure and consistent naming conventions for defined content across all guideline types will help ensure NICE content is classified and modelled correctly and to the same level of granularity.  In order to access full benefits of linked data in drawing together content across the institute this would need to be adopted universally i.e. not just for clinical, social care and public health, and would also need to incorporate the NCCs. |
|  | **1.2** Technology Appraisals  While the main focus of the project was guidelines, some related Technology Appraisals were included into the dataset to support the scenarios. Extracting the data from Technology Appraisals was also challenging. For the guidelines a set of recommendations were linked to a specific topic (sub-topic of the main guideline topic) along with accompanying evidence statements, discussion etc. However for Technology Appraisals there are only two “topics” – clinical effectiveness and cost effectiveness and any recommendations are relevant to both topics. | The current model may need to be modified for Technology Appraisals |
|  | **1.3** Accessibility of underlying evidence for Technology Appraisals  It proved very challenging to find underlying information on studies, search strategies and questions for Technology Appraisals as these form part of the manufacturer’s submission. However the discussions of the manufacturer’s submission referred to particular studies within their submission and it would have been useful to be able to follow these through with more transparency. No studies have been linked to this Technology Appraisals for this project as they were eventually found in a locked pdf document. | There are some implications for openness but this may depend on intellectual property issues with regard to manufacturers’ submissions.  Is there an opportunity to discuss with Manufacturers sharing more of this information? Explaining the direction NICE is taking with its content and explaining the usefulness of exposing more of this information? |
| 1. Observations on Guideline content | **2.1** Wording of recommendations and evidence statements.  Since the guidance is currently presented primarily in document format, many recommendations contain the wording “see recommendation above” or “see recommendation below”. Similarly, some evidence statements start with the words “ another study shows” which means that it cannot be understood in isolation without having read the previous evidence statement. This wording will only work in the context of a document with the recommendations or evidence statements in a particular order. As soon as any recommendations or evidence statements are extracted, the wording has no meaning. | Implications for modelling. For back catalogue recommendations will have to be presented in groups or the order modelled. For future recommendation wording should avoid this language. Technology needs to support alternative cross-referencing. |
|  | **2.2** Unique identifier for ‘entities’ e.g. recommendations and evidence statements  The numbering of recommendations is not consistent between different versions of a guideline (full guideline versus NICE guideline, for example). This made the extraction difficult and confusing since there is no unique identifier for a specific recommendation. It is made more confusing when a new piece of guidance “partially updates and replaces” another, as some recommendations will persist from one guideline to the next and some will be superseded, but the new guideline may reproduce the recommendations from the older guideline which are still valid and give them a new number, while they still technically exist with a different number in different guideline which has only been partially replaced.  Recommendations do not currently carry titles, only numbers. For the purpose of this exercise we created a title for each recommendation to find some way of distinguishing one recommendation from another within a grouping of recommendations. In some of the guidelines it proved challenging to capture what was unique to a particular recommendation since there was considerable repetition and overlap between separate recommendations. Similarly evidence statements have been assigned a title for this project but do not always have titles, in the guidance. | If each recommendation and evidence statement is to be modelled as an entity each unique recommendation will need a unique identifier. Currently there is no unique identifier for a recommendation or evidence statement, which could be a number or a unique title |
|  | **2.3** Evidence Strength  The system used for this varied widely between guidelines so was not included for most of the evidence statements. Most of the data attached a rating to a specific study. | A consistent system for assessing strength of evidence will allow analysis of whether the same study is rated consistently across various NICE guidelines.  (recognise that this is already a consideration of the methods working group and also wider groups such as DECIDE) |
|  | Studies linkage to evidence statements and recommendations  Studies proved straightforward to link to evidence statements, however sometimes they were listed in an appendix in a separate document and sometimes immediately after a grouping of evidence statements. For one of the guidelines (CG10) there appeared to be a few gaps where a study was mentioned in an evidence statement but could not be found in the list of references. Where a guideline had been “ partially updated and replaced” or “ superseded”, it was still sometimes the case that the references still had to be found in the appendix of a full guideline that had nominally been superseded (CG66/CG87). | Linking of study references to evidence statements and recommendations rather than the guideline as a whole would make the updating of guidance more straightforward, rather to hold reference details in a guideline which may have been superseded, even if the recommendation has not. |
|  | **2.3** Study unique identifier  Studies were listed within the context of a given guideline so it is likely that there is considerable duplication of studies referenced between guidelines. A workaround has been done to de-duplicate based on the textual content of the references | Using the PUBMED ID as a unique identifier for a study reference during the authoring process would enable analysis to be performed on studies referenced across all guidance.  It may also be beneficial to consider management and publishing of references using the BIBLIO ontology. |
|  | **2.4** Local practice collection  Shared Learning already had location, organisation and linked guideline as structured fields, which made it easier to extract the content. However this is currently not searchable. In addition, the shared learning ( local practice) only links at the guideline level, not at the recommendation level, although in reality no single case study attempts to implement the whole of a guideline. In extracting the data an attempt was made to link the local practice at recommendation level. Sometimes this was explicit within the text of the case study.  . | Modelling the location and organisation elements of the local practice collection and mapping local practice case studies to specific recommendations rather than to a full guideline would enable more targeted queries to be run by users and greater opportunities for linkage to other data sets.  Since this is not content produced initially by NICE it may be beneficial to revise the information that we ask people to submit and to control the methods by which they submit to ensure that they follow/align to our model(s) |
| 1. Observations on initial domain modelling exercise with NICE internal experts. | 3.1 Input from internal experts required for modelling NICE domain  Linked data requires an understanding and exposure of the relationships between concepts so there was a need to identify and validate both the key concepts in NICE guidance that would need to be modelled and the implicit relationships between those concepts.  An initial model rendered in Visio of the ‘NICE irreducible core’ concepts was validated with key internal experts and the relationships discussed. Drawing out concepts and articulating the exact nature of the relationships between those concepts requires a new way of thinking which can prove challenging to those who are used to viewing NICE content from a document-centric viewpoint. | Continued engagement with internal experts will be required to extend and refine the model. In order to access their time and expertise for modelling in a way in which they are unfamiliar, the potential benefits of linked data will need to be clearly articulated and reinforced at all levels. |
| 1. Observations on Ontology Modelling | **4.1** Challenges of ontology modelling  It can be challenging to decide whether to model something as a relationship or a class, or whether, for example “Topic” is an entity in its own right or a classification There is no one “correct” way to represent a domain in a model, but the way that is chosen will influence how the data can be queried and there are emerging standards for ontology modelling which encourage interoperability | Standards for ontology modelling will need exploring further and continued consultation with those organisations such as Cochrane with whom we may link data to ensure consistency. |
|  | **4.2** Defining the line between content, data and metadata  It has been challenging to model both the different concepts of the entities such as “recommendation”, “evidence statement” etc. and the content of these entities within the same model. The former are “NICE concepts” and the latter are universally understood concepts such as disease, drug, symptoms etc. which could be rendered by annotations or modelled in a more fundamental way. | The choice of annotating blocks of content vs modelling fundamental concepts and relationships needs to be taken case by case |
|  | **4.3** ‘As is’ versus ‘to be’ model  The model created for this project represents an incomplete but hybrid model of the “as is” and “to be” NICE domain. Currently in NICE it is not possible to link back from recommendations to specific evidence statements. Normally a group of evidence statements is presented alongside a discussion and group of recommendations, all under the heading of a specific sub-topic of the main guideline topic. Our Visio model models the Discussion concept “using” Evidence statements to “generate” a Recommendation. However, the data does not support this. For our scenario example on gastro-paresis, a close analysis of the data showed that all four evidence statements in one set (topic) were relevant to only one recommendation. The other two were supported only by the discussion. We have created an example for demonstrator to show how for this clear example, a gastro-paresis recommendation could link back through the discussion to specific evidence statements, but for most of the data it would be difficult to link specific evidence statements to specific recommendations. | Greater transparency may be achieved through a more direct linkage between recommendations and relevant evidence statements.  (recognise that this is already being discussed / addressed via the process working group) |
|  | **4.4** Evidence review and study model  The evidence review and study parts of the model have not been modelled in any great detail as the team is aware that Cochrane are doing a lot of work in this area and it may prove beneficial to wait and see what model they develop and see if it will work for NICE. | Continued discussion with Cochrane would be beneficial to understand and potentially influence their modelling to the benefit of NICE. |
|  | **4.5** Modelling People and Organisations  While it has not been done for this part of the project, there is a lot of value to be gained from modelling the people, the organisations and the roles involved in producing NICE guidance. A well-established ontology, FOAF, already exists that can be used to model these entities and the relationships between them, and doing this for NICE will be a key element of linking our internal data to external data. Most of what we have done in this project so far has been extracting and making internal links between NICE content, but to leverage true benefit this needs to join up with external linked data | Input from internal experts required to support appropriate modelling of people, stakeholders, organisations and groups. |
|  | **4.6** Audience  The concept of Audience, which is used for quality standards and public health recommendations has not yet been modelled for this project. It has been used in the guidance to suggest what specific actions a group of people need to take with regards to the associated guidance. Quality Standards use four categories of audience, while the public health guidelines have a more granular break down.  Including audience in the content precludes its use as metadata to find and target information by audience. However capturing and formally modelling audience metadata can be contentious as it is open to being misused/misunderstood and may preclude certain people from finding relevant information rather than helping. | Should audience be modelled using a standard controlled taxonomy that already exists within NICE or should we build into the model the fact that organisational change often means that duties shift between organisations and roles and that a task-based model may be more future-proof. |
|  | **4.7** Temporal modelling  Temporal modelling will be an important part of modelling NICE guidance to be able to analyse how and why guidance and evidence changes over time, but has not been attempted yet. It is not supported natively by RDF and will be relatively complex to model | Further Investigation required into various techniques for representing temporal aspects of data |
|  | **4.8** Entities included in the model   * Recommendations * Evidence Statements * Discussion * Study * Search Strategy * Topic * Guideline * Quality Standard * Quality Statement * Shared Learning * Question * Search Strategy * Outcomes Framework * Indicator * Measure   Entities not included in the model   * Research recommendations * Gaps in the evidence * Scope * Economic models * Implementation Tools * People * Organisations * Role * Audience | Not all these entities have been modelled fully and will still require input from internal experts to refine the model.  Priorities for future modelling need to be established |
|  | **4.9** Ontology Modelling Tools  Protégé[[10]](#footnote-10), the free open source ontology editor developed by Stanford University in collaboration with Manchester University and the de facto standard was used as the tool for modelling the NICE Ontology using the OWL 2, an ontology language for the semantic web. The full, desktop version of Protege was used rather than the pared- down web based version. The tool includes various reasoners which are used in complex ontologies to ensure that the ontology has been logically constructed and can be used to infer relationships which might otherwise be missed.  The tool continues to be actively developed and improved. | This tool is free for anyone to use |
|  | **4.10** Training to use Protégé  There are several online tutorials for using Protégé which were completed by the team and one team member completed a 2 day introductory tutorial at Manchester University. Doing the course, while not essential, did provide the confidence to tackle the initial ontology modelling and highlighted the thought processes required to build an ontology and the potential pitfalls inherent in the process. A particular feature was the familiarisation with the Open World Assumption used in the Semantic Web. This key concept is based on the assumption that what is not known to be true is simply unknown. This is in contrast to the closed world assumption of the database world where what is not known to be true is assumed to be false. However it could not cover all the modelling issues that are likely to arise in a complex ontology. | Although the tool is free some training is required to understand the concepts behind ontology modelling. For someone new to this, the Manchester University course may be useful as an introduction and is not expensive. |
|  | **4.11** Future ontological modelling  For this project modelling has been done at a relatively simple level, with minimal restrictions on classes and properties and without clear definitions and restrictions on what are the minimum requirements for something to be, for example, a recommendation or an evidence statement. However in future work, more sophisticated modelling will be required to specify the classes in more detail as without more restrictions being made explicit, the Open World Assumption means that the reasoner can return unexpected results. There is available a large ontology modelling online community who will offer their expertise to help with specific queries but a good understanding of the principles behind ontology modelling will be required to take this work forward. | The team has gained enough modelling expertise to produce a simple ontological model of the NICE domain in a short period.  Ontology modelling understanding will be required to support any future modelling work. |
|  | **4.12** Use of external ontologies  The decision was made to use some external ontologies such as PROV which has been designed to provide a set of classes, properties, and restrictions to represent provenance information, Open Annotation which specifies a framework for creating associations between related resources, and annotations and the W3C Content vocabulary[[11]](#footnote-11) to provide a framework to semantically represent textual content.  Further understanding of how to identifiy and assess existing ontologies and data sources will be important for any serious adoption of linked data. | Using external ontologies where appropriate is important for interoperability. |
| 1. Observations on Data | **5.1** Data modelled to date  The data extracted from guidance was all around the topic of Diabetes and related conditions as specified in the scenarios. This provided a large enough corpus to illustrate the principles of linked data as applied to internal content. The data included 6 guidelines, 2 technology appraisals, 1 quality standard, 3 outcomes frameworks, 3 local practice examples and 1 IPG. Separation of the content, creating relationships between the individuals within the identified classes and mapping it to the ontology schema has resulted in around 500,000 triples held in the triple store. | This is a small fraction of NICE content. Tackling the back catalogue will require time and resource. |
| 1. Observations on Technology | **6.1** Use of Stardog[[12]](#footnote-12) as triple store  Stardog was used as the graph database for this project. Stardog supports the rdf data model, Sparql queries and Owl 2. And was recommended by Manchester University as powerful enough to support the queries we need for this project, without the added complexity of OWLIM which is the tool used by the BBC. The performance of Stardog for the data we have there currently has been fine, but is untested at the 100m + triples level that is likely to be reached with all the guidance and provenance data that would need to be stored to fully support NICE content. Store performance depends heavily on the form of the data and nature of the queries and we would need to have a good estimate of the likely future data size to determine if this is likely to be a problem. | Stardog would require operational spike testing to assess suitability as a tool. It may not be preferable for any applications to use a triple store directly. The triple store would be offline and could then be used to generate content / search indexes and for when there is no other option but an ad-hoc query. |
|  | **6.2** Mapping data to the ontology  Keeping the code model synchronised with the ontologies proved difficult. The type provider created and implemented by the development team during the project helped with this. There are methods that can be used in Stardog to enforce individuals to conform to the ontology, but these are not portable between stores. Currently we are working with partial data and as a proof of concept, accuracy is not as critical as it would be in a production environment | Future development of bulk import should also use some verification queries .and use of hypermedia APIs would help prevent some of the spelling related issues encountered with ETL. |
| 1. Observations on querying data | **7.1** Querying the data using SPARQL  Querying the data successfully requires an understanding that all the NICE data is now represented in the form of triples. Given one piece of data from the triple, either or both of the other variables can be queried. SPARQL is a relatively simple query language to work with but it has a steep initial learning curve Other than those who are working closely with the model, it is unlikely that internal or external users would ever use actual SPARQL syntax and queries would need to be carried out as ready constructed queries choosing specific variables , hidden behind a user-friendly interface | A client facing application will need a user friendly interface on top of the underlying SPARQL queries |
|  | **7.2** Formulating successful queries  A good understanding of how the data has been modelled is essential to be able to formulate successful queries. The most progress in querying the data was made when an information architect with a good knowledge of the model and the data was sitting with a developer to work through SPARQL queries.  Only through the running of real queries on the data and viewing the results did flaws in the ontology come to light. Any required amendments to the model were made within Protégé, saved as an updated Owl file to be pulled into Stardog and the data re-imported. This could prove to be a time-consuming process. | An understanding of the ontology model is crucial to formulating good SPARQL queries on the underlying data. Knowing SPARQL on its own is not enough. |
| 1. Observations on Annotations and vocabularies | **8.1** Retrospective annotation of content  For this project the exercise has been in assigning concepts and applying tags retrospectively which is much more labour intensive than assigning terms at the time of authoring. To date NICE has not assigned terms at a more granular level than the guideline level. | Assigning terms and concept to content during the authoring process is more efficient than doing this retrospectively but will require the appropriate authoring tools. |
|  | 8.2 Concept Extraction using NLP tool  Textrazor, a Natural Language Processing tool was used to identify concepts within the textual content of the recommendations ,evidence statements, discussion and rationale with which to tag and match to concepts in Freebase [[13]](#endnote-5) , where possible, using the Open Annotation ontology to model the annotations. Freebase may not be the most appropriate concept store to link to, but has been used as a proof of concept for matching to standard concepts in vocabularies such as SNOMED CT[[14]](#endnote-6) and illustrates the principles of Linked Data where ontologies can be linked by identifying a concept in the home ontology by its URI in an external ontology. From Freebase it is possible to link to DrugBank, which contains information on contraindications and interactions for a specific drug that NICE has recommended for a specific disease. DrugBank holds a different set of information on drugs to that held by the BNF, but has been used as a proof of concept for how BNF data could be queried in the future if it were modelled in this way. | This proves that it is possible to link concepts discussed in NICE content to other data stores by using common URIs to refer to the content. |
|  | 8.3 Standard Vocabularies  No decisions have been made yet about which vocabularies/ontologies would be the most appropriate to link to. MESH is widely used in searching for evidence, but SNOMED CT[[15]](#footnote-13) is a likely front-runner for identifying medical concepts, given its support within the NHS and use in other relevant linked data projects such as Cochrane Linked Data project[[16]](#footnote-14)When researching these vocabularies it was found that not all the vocabularies NICE might be interested in using currently have an authorised owl version. Some versions have been created as an academic project by a third party -normally a university - which means that the version available online may be either inaccurate or not kept up to date. However there is a move from the official publishers of these vocabularies to produce official owl versions of them for use in ontology modelling. | Decisions required on preferred vocabularies for modelling concepts and tagging. |
|  | **8.4** Internal Vocabularies  Using standard, commonly understood concept terms is best practice for Linked Data, but if there was a need to annotate or tag NICE content with NICE specific terms from the NICE vocabulary pool, there would be a requirement to model the vocabulary using SKOS[[17]](#footnote-15), which provides a standard way to represent knowledge organization systems such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web. Only by modelling in this way can inference be applied on these concepts in the model. | Resource will be required to model any NICE vocabulary using SKOS |

1. <http://www.w3.org/TR/owl-features/> [↑](#footnote-ref-1)
2. <http://www.w3.org/TR/prov-o/> [↑](#footnote-ref-2)
3. <http://www.w3.org/ns/oa> [↑](#footnote-ref-3)
4. [↑](#footnote-ref-4)
5. <http://www.nlm.nih.gov/mesh/> [↑](#footnote-ref-5)
6. <http://www.drugbank.ca/> [↑](#footnote-ref-6)
7. <http://nicelinkeddata.wordpress.com/> [↑](#footnote-ref-7)
8. <http://xmlns.com/foaf/spec/> [↑](#footnote-ref-8)
9. <https://www.textrazor.com/> [↑](#footnote-ref-9)
10. <http://protege.stanford.edu/> [↑](#footnote-ref-10)
11. <http://www.w3.org/TR/Content-in-RDF10/> [↑](#footnote-ref-11)
12. <http://stardog.com/> [↑](#footnote-ref-12)
13. [↑](#endnote-ref-5)
14. [↑](#endnote-ref-6)
15. [↑](#footnote-ref-13)
16. <http://www.cochrane.org/community/development-projects/cochrane-linked-data-project> [↑](#footnote-ref-14)
17. <http://www.w3.org/2004/02/skos/> [↑](#footnote-ref-15)