MIM: A Minimum Information Model Vocabulary and Framework for Scientific Linked Data

Matthew Gamble, Carole Goble
School of Computer Science
University of Manchester, UK
firstname.lastname@cs.manchester.ac.uk

Graham Klyne, Jun Zhao
Image Bioinformatics Research Group
Department of Zoology
University of Oxford, UK
firstname.lastname@zoo.ox.ac.uk

***Abstract*—Linked Data holds great promise in the Life Sciences as a platform to enable an interoperable data commons, supporting new opportunities for discovery. Minimum Information Checklists have emerged within the Life Sciences as a means of standardising the reporting of experiments in an effort to increase the quality and reusability of the reported data. Existing tooling built around these checklists is aimed at supporting experimental scientists in the production of experiment reports that are compliant. It remains a challenge to quickly and easily assess an arbitrary set of data against these checklists. We present the MIM (Minimum Information Model) vocabulary and framework which aims to provide a practical, and scalable approach to describing and assessing Linked Data against minimum information checklists. The MIM framework aims to support three core activities: (1) publishing well described minimum information checklists in RDF as Linked Data; (2) publishing Linked Data against these checklists; and (3) validating existing "in the wild" Linked Data against a published checklist. We discuss the design considerations of the vocabulary and present its main classes. We demonstrate the utility of the framework with a checklist designed for the publishing of Chemical Structure Linked Data using data extracted from Wikipedia as an example.**

Keywords-Minimum Information Checklists; Scientific Linked Data; Data Reuse; Data Quality; Semantic Web;

I. INTRODUCTION & MOTIVATION

The rapid growth and wide adoption of the Linked Data approach [1] is underpinned by the openness of the Web platform. Using established web standards such as URIs and the Resource Description Framework (RDF), Linked Data provides a web-scale and open approach to data integration. Everybody can publish their data on this open Web, make replicates and host them at distributed locations on the web. This openness has particularly attracted the attention of the Life Sciences community. Consequently, an increasing volume of biological data has been made available in Linked Data format, for example: Bio2RDF [2], Chem2Bio2RDF [3], LinkedLifeData [4] and the recently founded and ambitious OpenPhacts [5] initiative.

This openness is a double-edged sword. On one hand, it drives a rapid growth of adoption and makes a large volume of data accessible on the web in a structured format. On the other hand, lack of governance and quality control has led to a Web of data of varied quality and trustworthiness [6].

A data quality issue caused by autonomous and distributed data publication is not new, nor is it unique to Linked Data [7]. A growing number of scientific data are being generated in a distributed manner, because rarely does a single research group have sufficient resources to generate data across a whole spectrum of varied complexities [8]. To enable 'big' science these distributed datasets must be gathered and integrated in order to create a big picture about what is known or what can be done. In the Life Sciences, the Biosharing initiative [9] is driving efforts to control the quality of the data published by a diverse range of research groups by gathering together and coordinating reporting standards to which data submitters must comply. Several classes of interoperable reporting standards are combined: reporting frameworks such as the ISA (Investigation, Study, Assay) framework [10]; data formats, controlled vocabularies, and Minimum Information Checklists [11]. Minimum Information Checklists (MICs) define a minimum list of information and attributes that must be included in the submitted data, and sometimes, the format in which the data is reported. These checklists cover a wide range of types of biological investigation with some 60+ MICs currently listed for structuring and curating data by Biosharing. The Minimum Information for Biological and Biomedical Investigations (MIBBI) project [11] in particular has highlighted the important role of MICs in the reporting of biological investigations. MIBBI is primarily focused on the 'Omics, where experiments are characterized by high volumes of output data with a significant potential for reuse. The integration of quality control in the process of making data accessible has led to the creation of a number of respected 'Omics databases, such as the ArrayExpress database that is regulated by the MIAME and MINSEQE MICs. An extract from an example MIC for bioactive chemical compound data is given in Fig. 1.

This research is partially supported by the Wf4Ever project (http://www.wf4ever-project.org), Project 270129 funded under EU FP7 Digital Libraries and Digital Preservation (ICT-2009.4.1).

This research has also received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement number 115191, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007- 2013) and EFPIA companies' in kind contribution

10.1109/eScience.2012.6404489

1. MIABE

1.1. MIABE-Compound

1.1.1. Molecule properties

1.1.1.1. Primary name (+ synonyms)[1]

1.1.1.2. Molecule type[2]

1.1.1.3. Chemical IUPAC Name

1.1.1.4. Chemical structure[3]

1.1.1.5. InChI String and/or Key[4]

1.1.1.6. Chemical salt[5]

1.1.1.7. Prodrugs[6]

Figure 1. Extract from the Minimum Information about a Bioactive Entity (MIABE) checklist. The checklist currently details 48 reporting requirements for bioactive compounds.

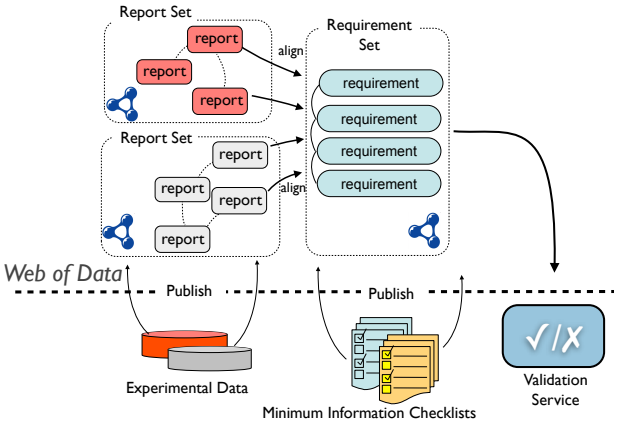

Figure 2. Reporting and requirement structures used to align MICs with scientific Linked Data.

Checklists currently exist in a number of formats (PDFs, Excel spreadsheets, XML Schema definitions) but not RDF. By making MICs available for Linked Data publishers, we would be in a better position to assess and increase the quality of scientific data on the Web of Data. We look to achieve this by supporting three core activities using MICs: (1) Supporting authorities, data providers and the community in publishing well structured checklists describing the minimum set of information required when publishing a particular class of data; (2) Supporting individuals, data creators and scientists in publishing Linked Data against the MIC; and (3) Supporting data consumers in assessing existing data "in the wild" against a MIC. The scope of our approach, and minimum information checklists in general, is to validate the reporting of requirements, not the correctness of the reported information. However, the validation of meta-data completeness is an important first step to increasing the quality of scientific information on the Web of Data.

A collaboration with the Royal Society of Chemistry (RSC) as part of the OpenPHACTS project has presented us with the challenge of addressing the quality of published chemical compound data. It is now community consensus in Chemistry that a good quality description about a chemical compound requires that it must provide an InChI (IUPAC International Chemical Identifier). This is a requirement that has been subsequently captured in the MIABE (Minimum Information About a Bioactive Entity) checklist [12]. The ChemSpider database [13], hosted and managed by the RSC, will not allow entries that do not provide an InChI identifier. However, given the open nature of the Web of Data we lose the ability to enforce this requirement. When publishing compound data into the Linked Data web an individual may discard this InChI. By expressing in a checklist that a unique InChI is a minimal requirement for information about a chemical compound to be complete, this flaw in the data can be more readily detected. If the data publisher aligns their Linked Data along with the MIC that they wish to be compliant with, this intended compliance is then carried along with the data. Data consumers will be able to make a better interpretation about the quality of the data, automatically check it and integrate and compare it with greater confidence. Moreover, given a well-structured representation of our checklist we are no longer restricted to data that was explicitly published against a particular MIC. In addition we can attempt to align and map existing data to the checklist and make an assessment of its compliance.

We identify three components for a MIC framework governing quality control of published Linked Data (Fig. 2): a *requirements description* that enumerates the requirements to be satisfied; a *report description* to align elements of the published data with requirements (both manually and automatically); and a *validation* to determine a level of conformance of these reports with the requirements of a MIC. We must do this in a way that copes with the wide diversity of MICs and data in the field. Our contributions are:

1. The Minimum Information Model (MIM) Vocabulary, a meta-modeling vocabulary used in three ways:
   - To describe a MIC that is specific to some class of data (e.g. MIABE for Bioactive compound data) with the aim of having a library of RDF encoded checklists as community resources for data of various types that can be referred to and used by anybody;
   - To annotate RDF data (e.g. data reporting a particular bioactive entity) as reports of requirements, and to aggregate these reports together into coherent sets that claim to satisfy a MIC; and
   - To express a level of conformance of some aggregation of reports with the requirements of a MIC.

2. A prototype implementation of a framework that combines requirements reported using the MIM vocabulary to calculate an assessment of conformance to a MIC; and

3. A case study demonstrating the viability and utility of our approach to evaluating data in the wild. Applying our framework we assess the completeness of records from a Linked Data extraction of the detailed chemical compound data available on Wikipedia.

Before presenting the results of our case study we provide a description of our MIM Vocabulary. Without sufficient space to detail all features we instead highlight core features and design decisions of the vocabulary and supporting framework. We illustrate throughout with examples drawn from a

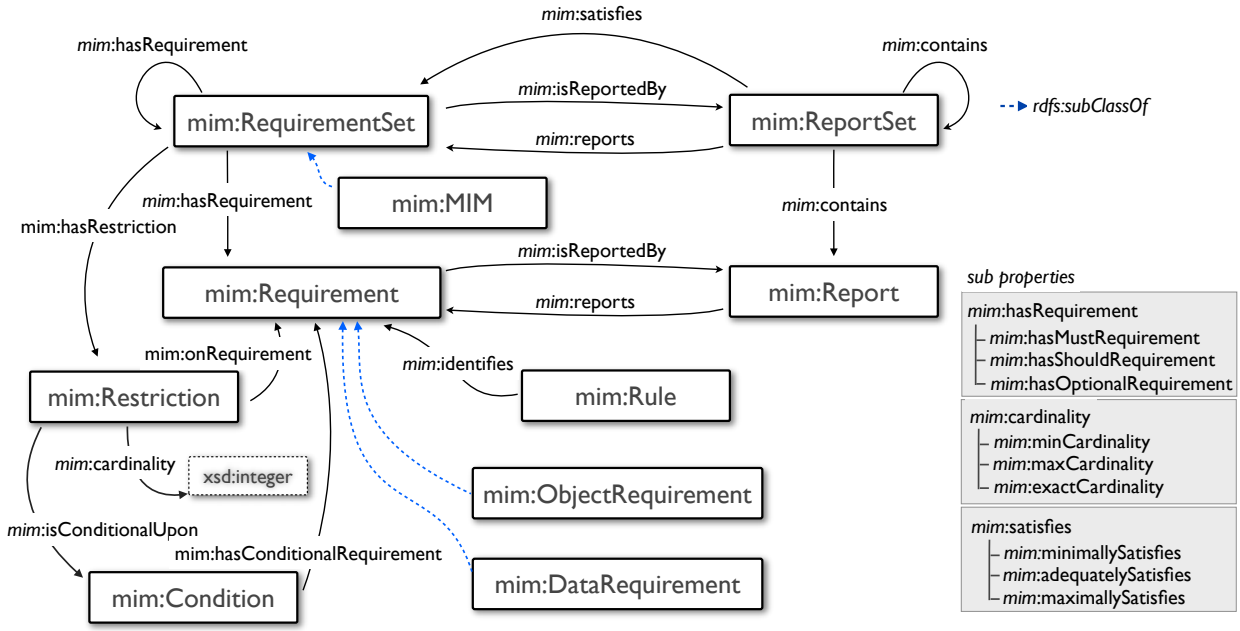

Figure 3. The core classes and properties of the Minimum Information Model (MIM) vocabulary

checklist, CHEMIM, developed for our case study. The CHEMIM checklist details 13 meta-data reporting requirements for publishing chemical compound data on the Web of Data.

II. THE MIM VOCABULARY

The MIM Vocabulary is a meta-modeling vocabulary intended to allow the encoding of arbitrary MICs. The MIM vocabulary itself has been developed as an OWL DL ontology (available at http://purl.org/net/mim/ns). Fig. 3 presents the core classes of the vocabulary and the relationships between them. The main classes can be separated into two areas of concern, *checklist description* and *report description*.

A. Describing A Checklist

The construction of a MIC serves to specify the minimal set of data to provide when reporting a particular class of data, or type of experiment. Fig. 4 details an extract from our example MIC encoded in the MIM vocabulary.

Specifically the metadata requirements captured in the extract detailed are:

- The compound report **must** provide an InChI and a Simplified Molecular Input Line Entry System (SMILES) value.
- The compound report **should** contain a ChemSpider ID, PubChem ID, an Image, and an International Union of Pure and Applied Chemistry (IUPAC) Name.
- The compound report **may** also report some synonym values.

There are a core-set of common features used in MICs which allow us to propose a domain-agnostic MIM meta-model suitable for describing the above requirements, as well as many existing checklists.

Requirements Requirements express the need to report an individual element of data. In our example lines 37-38 detail the required reporting of a SMILES value. This requirement is specifically a `DataRequirement` where the expected report is a text or numerical value i.e. an RDF literal value.

Requirement Sets Across MICs currently in use there is a significant level of overlap and reuse of requirements [11]. As a result the MIBBI project promotes a modular approach to MIC construction. In our CHEMIM checklist we specify the `Identifiers` requirement set which gathers together the InChI, SMILES, PubChem, and ChemSpider ID requirements. This Identifiers set is then potentially reusable by another checklist.

Requirement Levels A prevalent feature of MICs is the indication of a level of requirement. A significant number of reporting guidelines are explicit in their use of the terms from RFC 2119 [14], namely: *must*, *should*, and *optional* (e.g. MIARE [15], and MIFlowCyt [16]). We have adopted these requirement levels into our vocabulary. For example the `Identifiers` requirement set states that `InChI` is a must requirement through the use of `mim:hasMustRequirement`. These requirement levels are subsequently taken into account when validating how well a checklist has been satisfied.

Cardinality Restrictions There is often a need to specify the numbers of a particular report permitted by a requirement set. For example a chemical compound should have one and only one SMILES value. Within our `Identifiers` requirement set specification, the restriction on lines 22-26 serves to specify that if reporting this requirement set you must report exactly one SMILES value.

Vocabulary Restrictions In the interests of interoperability the Life Sciences have a significant investment in the use of controlled vocabularies and ontologies [17][18]. Vocabulary constraints are therefore often stated explicitly in many of the existing checklists. The RightField tool [19] for example

```
1  @prefix mim: <http://purl.org/net/mim/ns#> .
2  @prefix :    <http://purl.org/net/chembox/
       chemmim#> .
3
4  :MIM  rdf:type mim:MIM ;
5    mim:hasMustRequirement
6            :Identifiers, :Properties ;
7    mim:hasOptionalRequirement
8            :Synonym ;
9    mim:hasRestriction
10       [ mim:exactCardinality
11                 1 ;
12         mim:onRequirement :Identfiers,
13                 :Properties
14       ] ;
15   mim:hasShouldRequirement
16           :IUPACName , :Image .
17
18 :Identifiers
19   rdf:type mim:RequirementSet ;
20   mim:hasMustRequirement
21           :InChI , :SMILES ;
22   mim:hasRestriction
23       [ mim:exactCardinality
24                   1 ;
25         mim:onRequirement :InChI , :SMILES
26       ] ;
27   mim:hasShouldRequirement
28         :ChemSpider , :PubChem .
29
30 :InChI
31   rdf:type mim:DataRequirement ;
32   mim:hasRestriction
33       [ mim:onSelf "true"^^xsd:boolean ;
34          mim:type xsd:string
35       ] .
36
37 :SMILES
38   rdf:type mim:DataRequirement .
39
40 :Image
41   rdf:type mim:ObjectRequirement .
42   mim:hasRestriction
43       [ mim:onSelf "true"^^xsd:boolean ;
44         mim:instanceOf foaf:Image
45       ] .
```

Figure 4. CHEMIM (extract) described using the MIM vocabulary. The checklist prescribes 13 metadata requirements for the reporting of chemical compound data.

captures this behaviour, enabling the integration of vocabulary restrictions into Excel spreadsheet representations of MICs. The MIM Vocabulary similarly allows for the description of such restrictions (e.g. line 44).

These core vocabulary features support the description of many of the requirements of existing checklists as machine readable RDF. These checklist descriptions can then be published in the Web of Data. Once published each requirement and requirement set becomes a uniquely identifiable Linked Data resource that data publishers may report against.

B. Reporting Against a Checklist

The objective when reporting against a MIC is to align source data with the checklist, making claims about which resources in the source data report the requirements specified in the MIC. Different Linked Data publishers often have alternative ways of expressing information that satisfies the same requirements. We therefore need a vocabulary that serves to annotate and align data to requirements that they claim to

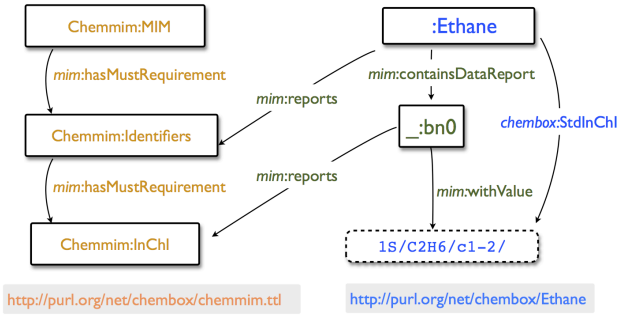

Figure 5. An example showing how a report of the InChI value in http://purl.org/net/chembox/Ethane.ttl is aligned with the CHEMMIM checklist.

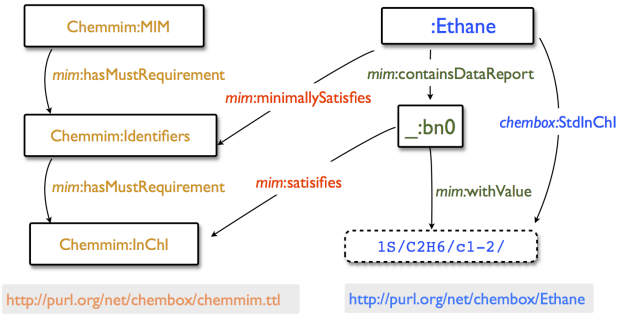

Figure 6. Example from Fig 5. subsequently annotated to show requirement satisfaction, using `mim:satisfies` and `mim:minimallySatisfies`.

report that is agnostic to its representation. For this we identify the need for the following:

Reports In RDF data we can identify two types of report that we need to annotate and align with the checklist: *Data Reports* - where the reporting data is an RDF literal value and *Object Reports* - where the reporting data is an RDF resource, uniquely identifiable by a URI.

Report Sets A report set is a collection of reports that claim to report a requirement set. In a given set of RDF data there may be multiple instances of a chemical compound, complete with all of their meta-data. To subsequently evaluate completeness and cardinality constraints it necessary to identify coherent report sets and align them with a corresponding requirement set in our checklist.

As a concrete example consider the RDF presented in Fig. 5. Here we have a report of an InChI for the resource `Ethane` aligned with our checklist. The identifier itself is presented as a literal value. This is aligned by constructing the `DataReport` resource `_:b0`. The `mim:reports` property indicates the requirement being reported and the specific literal value using the `mim:withValue` property. The resource `Ethane` is annotated as a report set, containing the report `_:b0` and reporting the `Identifiers` requirement set.

Report Generating Rules The task of aligning data with a checklist is only feasible by hand annotation for small-scale data. Scientific Linked Data is typically large scale and inconsistently represented across data sets [20]. To retrospectively align existing data with MICs we need to automate this process. The vocabulary term `mim:Rule` serves as a mechanism to align rules for report generation with particular requirements in the checklist. The vocabulary is agnostic to the particular implementation of the rule

```
CONSTRUCT {
    ?x mim:containsDataReport _:b0 .
    _:b0 mim:reports :InChI .
    _:b0 mim:withValue ?y .
}
WHERE {
    ?x chembox:StdInChI  ?y .
}
```

Figure 7. An example of a SPARQL query used as a report-generating rule. The query generates the report structure in Fig 5.

mechanism. The only requirement is that the rules can be uniquely identified with a URI.

C. Checklist Satisfaction

Given a set of RDF data aligned with a MIC our goal is to validate how well that data satisfies the checklist. MIC satisfaction can be divided into two broad concerns:

Individual Requirement Satisfaction: Given a claim of the type `_:b0 mim:reports chemmim:InChI`, where `chemmim:InChI` is a *requirement*; we say that `_:b0` *satisfies* the requirement if it meets any type constraints specified. In this case our checklist specifies the report should be a text value (of type `xsd:string`).

Requirement Set Satisfaction: Given a claim of the type `Ethane mim:reports chemmim:Identifiers`, where `chemmim:Identifiers` is a *requirement set*; we say that `Ethane` *satisfies* the report set if (1) it contains reports that satisfy `chemmim:Identifiers` requirements (e.g. `_:b0`) and (2) it meets any cardinality restrictions defined by `chemmim:Identifiers`. We recall that a requirement set may specify one of three levels of requirement (must, should, optional). As a result we have defined three progressively more complete levels of requirement set satisfaction:

Minimally Satisfies: The report set satisfies all *must* requirements.

Adequately Satisfies: The report set satisfies all *must* and all *should* requirements.

Maximally Satisfies: The report set satisfies *all* requirements.

If our annotated source data meets the above conditions then we can indicate as such using vocabulary terms. Fig. 6 details the same data as Fig. 5 complete with the subsequent annotations to indicate requirement and requirement set satisfaction.

III. Implementation

To validate our approach and support our subsequent case study we developed a prototype implementation of our MIM framework[1]. The prototype has been developed as a Java Web Service and provides two key functionalities: (1) Automatically identifying reports in source data and aligning them with a checklist; and (2) Validating data aligned with a checklist, that is, how well the data satisfies the checklist.

To realize both the report-generating rules and the rules governing our MIM satisfaction semantics we have used the SPARQL Inferencing Notion (SPIN) [21]. SPIN is a standard in submission for representing SPARQL (SPARQL Protocol and Query Language) rules and constraints over RDF data. The standard defines a serialization that allows the definition and publication of SPARQL queries as RDF. These queries can then be applied as rules to make inferences over RDF data and perform integrity constraint checking. Therefore, it provides the exact mechanism that we need to encode our report generating rules and MIC satisfaction semantics. Tooling already built around SPIN[2] makes it easy to support the definition of report-generating rules. Using SPIN we can specify a rule that generates the report of a `chemmim:InChI` for any data in the same format as Fig. 5. The SPARQL rule to generate this report is detailed in Fig 7. By serializing and publishing this rule as RDF others can reuse or extend it. Indeed the specific implementation of our MIM satisfaction semantics (validating requirement and requirement set satisfaction) is a collection of SPIN rules available on the web at http://purl.org/net/mim/mimspin.

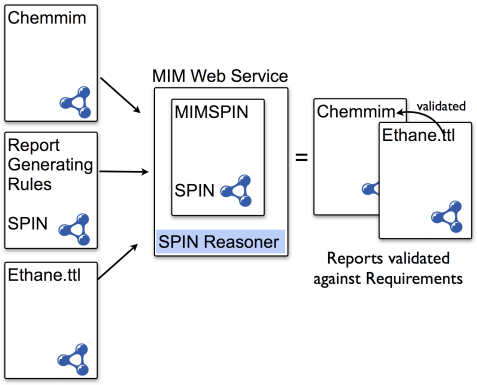

Figure 8. Data flow depicting the process of validating a data against a checklist using the MIM validation web service.

The MIM web service builds upon the TOPSPIN API[3], an Apache Jena based implementation of a reasoner that can process SPIN rules and generate the necessary inferences over our data. Fig. 8 details the process of calling the web service to validate a set of data against a checklist. Calling the web service requires passing it three arguments: a URL for the checklist defined using the MIM vocabulary, a URL for the data, and a URL for a set of SPIN encoded report-generating rules for that data set. The web service hands these to the SPIN reasoner which runs the supplied rules over the data and returns the newly inferred triples aligning and validating the data with the checklist.

IV. A WikiProject Chemicals Case Study

To demonstrate our MIM framework we have chosen to perform large-scale MIC assessment of the chemical compound data already published in Wikipedia. Wikipedia is steadily becoming a valuable reference resource for chemical compound data [22]. However, as an open community data resource the data available is of varying quality and completeness. Since 2005 Members of the Wikipedia task force *WikiProject Chemicals* [26] have made attempts to assess the quality of the chemical compound entries by hand. As of

[1] http://github.com/matthewgamble/mim-ws

[2] http://www.topquadrant.com/

[3] http://topbraid.org/spin/api

writing there still remain several thousand entries un-assessed. We therefore choose to demonstrate the utility of the MIM framework by assessing the completeness and conformance of the data available on these pages to a MIC developed for chemical compound reporting. This data is currently presented as property-value pairs in the form of *Infoboxes* and as such it is possible to generate a Linked Data extraction for each chemical compound entry. The members of WikiProject Chemicals are concerned with the completeness and quality of whole entries (article text and Infobox data). Our MIC assessment serves as a first level of quality assessment by checking the compound data in the Infoboxes for meta-data completeness.

There is an existing and widely used Linked Data extraction of Wikipedia's Infoboxes in the form of DBpedia [23]. We performed our own extraction for two reasons. Firstly, the current DBpedia extraction of chemistry Infobox data is very limited and insufficient to test our approach (and does not sufficiently reflect the chemical data available in Wikipedia). Secondly, by performing our own extraction we gain insight into how the entire process of generating and publishing Linked Data impacts our MIC assessment. Scientific data sources are rarely created initially as Linked Data resources; instead a Linked Data representation is typically generated from a primary existing source [20]. As such our case study is representative of much of the existing scientific data available on the Web of Data. We developed a tool[1] to extract data from Wikipedia's Infoboxes, building upon a Java-based MediaWiki API, the Java Wikipedia Library [24]. Using this tool we extracted Linked Data representations of 7572 chemical compound pages from Wikipedia (e.g. http://purl.org/net/chembox/Ethane.ttl). The full Linked Data Chembox extraction totals 376,282 RDF triples[2]. Our extraction enhances the current DBpedia extraction so in adherence to Linked Data principles, we have linked our data back to the existing DBpedia resources.

To perform a MIC assessment over our data we have defined a MIC, CHEMMIM[3] using the MIM Vocabulary. Our checklist is designed to be representative of the meta-data typically required when reporting chemical compound data. The chemistry community on Wikipedia currently has a guideline similar to a MIC in the form of an Infobox template, Chembox. The 'simple' Chembox template[4] details a set of 15 meta-data properties that the community recommends for inclusion when creating a chemical compound article. Taking a number of shared elements from both the Chembox template and the MIABE checklist (specifically the molecule properties section) the full resulting CHEMMIM checklist defines 13 reporting requirements:

Must: Molecular Formula, InChI, and SMILES.

Should: Melting point, Molar mass, ChemSpider ID, PubChem ID, IUPAC name, and Image;

Optional: Synonyms and Solubility.

The checklist also groups some of these requirements into two reusable requirement sets: *Identifiers* – InChI, SMILES, ChemSpider ID, PubChem ID; and *Properties* – Molecular Formula, Molar Mass, Melting Point and Solubility. The incentive for defining these requirement sets is that upon validation we gain a MIM completeness assessment (minimally, adequately, maximally) for the requirement sets, as well as the checklist as a whole.

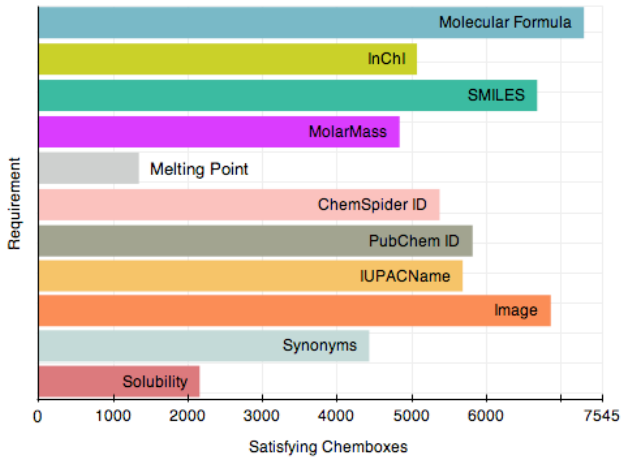

Figure 9. Individual requirement satisfaction across all Chembox instances.

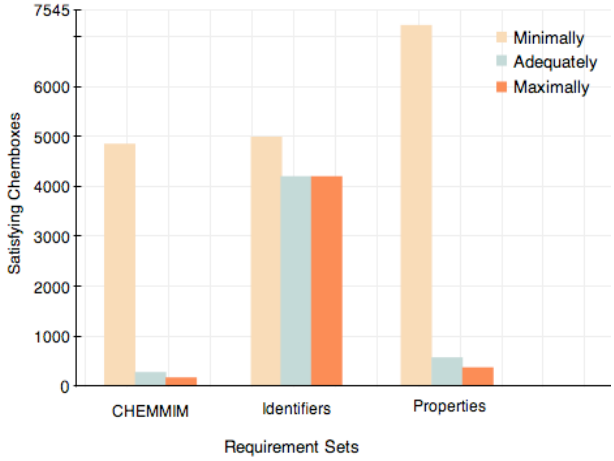

Figure 10. Requirement set satisfaction across all Chembox instances.

Along with the checklist we have also defined a set of 15 report-generating rules[5] in the SPIN format to annotate and align reports in the Chembox Linked Data with the CHEMMIM checklist.

Using our prototype MIM validation web service we have validated each of the 7572 Chembox extractions against our CHEMMIM checklist. The validation process generated a further 808,420 triples which align the data with the checklist and detail requirement satisfaction. With this validated data in a well-structured form we can query it and start to ask questions about how well chemical compound data is being reported across Wikipedia.

The graphs in Fig. 10 and Fig. 11 present two views over the data. Fig. 10 shows how each individual requirement in our checklist is satisfied across the entire Chembox data set. The

[3] https://github.com/matthewgamble/chembox

[4] http://purl.org/net/chembox/

[5] http://purl.org/net/chembox/chemmim

[6] http://en.wikipedia.org/wiki/Template:Chembox

[7] http://purl.org/net/chembox/chemboxspin

most well satisfied reporting requirement is Molecular Formula with 7263 (96%) of Chemboxes satisfying the requirement. In contrast to this our results show that Melting Point data is particularly poorly reported across Wikipedia with only 18% (1343) of Chemboxes satisfying the requirement. This finding is validated by recent activity within the online chemistry community to improve the availability of open melting point data [25].

Fig. 11 shows how well the requirement sets; Identifiers, Properties and the MIC itself are satisfied. Currently 64% (4863) of Chemboxes minimally satisfy the CHEMMIM by satisfying all of the must requirements. The number of Chemboxes going beyond minimal satisfaction drops off significantly with only 274 adequately (reporting all must and should requirements) and 168 maximally (reporting all requirements) satisfying the checklist. The WikiProject Chemicals hand-classified assessment can be seen to be in agreement with this finding [26]. On a 4 point scale of Stub, Start, B, and A, the task force currently rate 4413 articles as Stub and only 16 as A. Though the remit of their assessments differs, this would suggest a similar discrepancy between the number of top quality and minimal quality articles. Our MIC assessment however provides us with a more detailed view. For our requirement sets our results reveal that whilst Wikipedia may be a poor source for the particular chemical properties we have defined, it is a relatively good source for chemical identifiers, with 56% (4207) of Chemboxes *maximally* satisfying the Identifiers requirement set. This more detailed view of the data afforded by our MIC assessment is lost in the WikiProject chemicals evaluation.

V. Related Work

There has been some previous effort to harmonize and structure the representation of MICs beyond their traditional flat text representation.

The MIBBI project is the effort most closely aligned with ours. The project currently hosts a web based tool MICheckout [27]. This tool allows the user to download existing checklists (such as MIABE) or compile their own custom checklists by selecting a number of reusable requirement sets. These MICs can then be exported as HTML, XML Schema definitions, tab-delimited files, and MediaWiki templates.

The RightField tool [19] provides the ability to specify MICs as Excel spreadsheets. In particular the RightField tool allows the checklist creator to restrict elements of a spreadsheet to particular ontological terms. This ensures that experimental reports subsequently created using the spreadsheet are compliant to a particular data format.

In the broader effort to improve the quality and interoperability of Life Sciences data, the ISA (Investigation, Study, Assay) framework and supporting tooling [28] is gaining significant adoption. In a move towards an 'ISA commons', the ISA framework relies upon data producers conforming to common metadata categories – Investigation, Study and Assay. Central to the ISA ecosystem is the ISA-Tab format. ISA-Tab is a hierarchical tab-delimited template that details minimum reporting requirements whilst ensuring data is captured in the ISA format.

These current efforts are focused on building tools and infrastructure that ensure the production of MIC compliant data. We believe we are the first to address the challenge of assessing data published "in the wild" for MIC compliance.

There has been specific effort to understand the generation and integration of scientific Linked Data. The authors in [20] have reviewed recent efforts of the Health Care and Life Sciences Special Interests Group (HCLS IG) in publishing a number of data sets as Linked Data resources. In doing so the authors make a series of recommendations to improve the quality and utility of scientific Linked Data.

Despite these related efforts to the best of our knowledge there has been no previous effort to develop a meta-model to enable the encoding of arbitrary MICs in RDF suitable for the Web of Data. The MIM vocabulary introduces a number of terms that are similar to the existing core OWL vocabulary (e.g. for cardinality restrictions). The MIC assessment assumes a closed world semantics i.e. if the data is not present then it is assumed not to exist. This differs from the typically assumed semantics of OWL constraints. We have chosen to introduce our own terms to make this distinction clear and to allow MIM descriptions to co-exist with existing OWL without confusion.

VI. Discussion and Conclusions

In this paper we have presented our work on bringing the utility of minimum information checklists to the web of scientific Linked Data. Fig. 10 details the final MIC assessment of what was an iterative process. For 35 (approx. 0.5%) of the chemical compounds, we manually derived a ground truth for the original source data via a consensus across multiples data sources. For these compounds we established which requirements were reported, and therefore how well they satisfied the checklist. Any anomalies triggered a process of fault-finding. These anomalies occur due to a number of reasons: (1) inadequate mapping rules; (2) poor extraction from the source data; or (3) the data is genuinely missing from the source data. Examining the RDF generated for our ground truth resources we were able to establish for example whether the data was present but failed to be mapped, or absent and failed to be extracted. Following this process we found the MIC assessment particularly useful in highlighting errors in the RDF generation process. Fault-finding was a non-trivial task with scope for future work to support and improve this process. Our case study also served to highlight that whilst we gain an indication of the completeness and quality of the original data source, it is the Linked Data extraction itself for which we are making the true MIC assessment.

The ability to perform a large-scale assessment of scientific Linked Data provides a number of benefits to data producers, consumers, providers and even the developers of checklists themselves. For the maintainers of community data resources such as WikiProject chemicals, large scale MIC assessment can be used to suggest where efforts would be best placed to improve the resource. Alternatively data consumers are presented with the opportunity to base their source selection on which source better satisfies the MIC requirements they are interested in e.g. Wikipedia and chemical compound identifiers.

The development of a MIC is a difficult process, checklists aim to fulfill the criteria of sufficiency and practicability i.e. not be so burdensome as to prohibit use [12]. A large-scale

analysis can provide feedback to the developers of checklists and suggest where they may be falling short of these criteria.

In future work we look to further validate our MIM framework, encoding a broader range of checklists and incorporating a wider array of the growing number of scientific Linked Data sources. A valuable first step would be to support the ever-growing ISA ecosystem.

Working with the Web of Data we are not necessarily constrained to one data source in order to satisfy a MIC. Our approach as it stands may make use of federated SPARQL queries - where we can address multiple data sets in one query. This is particularly applicable to Life Science data, where related components of a study may be scattered across disparate resources [9]. Incorporating data for multiple data sources raises the related issues of provenance and trust. We see the incorporation of detailed provenance of the report generation and MIC validation processes as a valuable avenue for future work for a number of reasons. *Provenance and versioning* - when revisiting data that claims to satisfy a MIC we wish to understand how it was determined that the data is compliant. *Provenance to aid fault-finding* – having detailed lineage about how reports were generated can aid in the fault-finding process when developing RDF generation and report-generating mappings. *Provenance for attribution* – if one were to subsequently use data aligned with a checklist in a further study, detailed provenance about where and when that MIC assessment was performed is crucial to give correct attribution.

Using our MIM framework we have successfully evaluated the chemical compound data available in Wikipedia against a MIC. As an increasing number of scientific data sets look to take advantage of the Web of Data, the need for techniques to aid in understanding and improving the quality of that data also increases. We believe the MIM Vocabulary and framework will be a great help in this effort.

ACKNOWLEDGMENT S

We would like to thank Dr. Antony Williams of the Royal Society of Chemistry for his valuable support. We are also grateful to all of the members of the Wf4Ever project.

REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data–the story so far," *International Journal On Semantic Web and Information Systems*, 2009.

[2] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge Systems.," *Journal of biomedical informatics*, vol. 41, no. 5, pp. 706-16, Oct. 2008.

[3] B. Chen et al., "Chem2Bio2RDF: a Semantic Framework for Linking and Data Mining Chemogenomic and Systems Chemical Biology Data.," *BMC bioinformatics*, vol. 11, p. 255, Jan. 2010.

[4] V. Momtchev, D. Peychev, T. Primov, and G. Georgiev, "Expanding the Pathway and Interaction Knowledge in Linked Life Data," *ontotextcom*, 2009.

[5] A. J. Williams et al., "Open PHACTS: Semantic Interoperability for Drug Discovery.," *Drug discovery today*, Jun. 2012. Available: http://dx.doi.org/10.1016/j.drudis.2012.05.016

[6] S. Schlobach and C. a. Knoblock, "Dealing with the Messiness of the Web of Data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 14, p. 1, Jul. 2012.

[7] A. J. Williams and S. Ekins, "A Quality Alert and Call for Improved Curation of Public Chemistry Databases," *Drug Discovery Today*, vol. 16, no. 17–18, pp. 747-750, Jul. 2011.

[8] D. Field et al., "Megascience. 'Omics data sharing.," *Science (New York, N.Y.)*, vol. 326, no. 5950, pp. 234-6, Oct. 2009.

[9] S.A. Sansone et al., "Toward Interoperable Bioscience Data," *Nature Genetics*, vol. 44, no. 2, pp. 121-126, Jan. 2012.

[10] S.A. Sansone et al., "The first RSBI (ISA-TAB) workshop: 'can a simple format work for complex studies?'.," *Omics : a journal of integrative biology*, vol. 12, no. 2, pp. 143-9, Jun. 2008.

[11] C. F. Taylor et al., "Promoting Coherent Minimum Reporting Guidelines for Biological and Biomedical Investigations: the MIBBI Project.," *Nature biotechnology*, vol. 26, no. 8, pp. 889-96, Aug. 2008.

[12] S. Orchard et al., "Minimum Information About a Bioactive Entity (MIABE).," *Nature reviews. Drug discovery*, vol. 10, no. 9, pp. 661-9, Sep. 2011.

[13] H. E. Pence and A. Williams, "ChemSpider: An Online Chemical Information Resource," *Journal of Chemical Education*, vol. 87, no. 11, pp. 1123-1124, Nov. 2010.

[14] S. Bradner, "IETF RFC 2119:Key words for use in RFCs to Indicate Requirement Levels." [Online]. Available: http://www.ietf.org/rfc/rfc2119.txt. [Accessed: 17-Jul-2012].

[15] "MIARE: Minimum Information About an RNAi Experiment." [Online]. [Accessed: 17-Jul-2012]. Available: http://miare.sourceforge.net/MIAREReportingGuidelines.

[16] J. a Lee et al., "MIFlowCyt: the minimum information about a Flow Cytometry Experiment.," *Cytometry. Part A : the journal of the International Society for Analytical Cytology*, vol. 73, no. 10, pp. 926-30, Oct. 2008.

[17] K. a Spackman, K. E. Campbell, and R. a Côté, "SNOMED RT: a reference terminology for health care.," *Proceedings : American Medical Informatics Association Fall Symposium*, pp. 640-4, Jan. 1997.

[18] B. Smith et al., "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.," *Nature biotechnology*, vol. 25, no. 11, pp. 1251-5, Nov. 2007.

[19] K. Wolstencroft et al., "RightField: Embedding Ontology Annotation in Spreadsheets.," *Bioinformatics (Oxford, England)*, pp. 1-2, May 2011.

[20] M. S. Marshall et al., "Emerging Practices For Mapping and Linking Life Sciences Data using RDF — A case series," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 14, pp. 2-13, Jul. 2012.

[21] H. Knublauch, "SPIN - Modeling Vocabulary. W3C Member Submission." [Online]. Available: http://www.w3.org/Submission/spin-modeling/.

[22] P. Murray-Rust, "Chemistry for everyone," *Nature*, vol. 451, no. 7179, pp. 648-651, 2008.

[23] C. Bizer et al., "DBpedia - A Crystallization Point for the Web of Data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 3, pp. 154-165, Sep. 2009.

[24] T. Zesch, C. Müller, and I. Gurevych, "Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary," in *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 2008.

[25] J.-Claude Bradley, A. Lang, A. Williams, and E. Curtin, "ONS Open Melting Point Collection," *Nature Precedings*, pp. 1-699, Aug. 2011.

[26] "WikiProject Chemicals Assessment." [Online]. Available: http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Chemicals. [Accessed: 10-Jul-2012].

[27] C. Kettner et al., "Meeting Report from the Second 'Minimum Information for Biological and Biomedical Investigations' (MIBBI) workshop.," *Standards in genomic sciences*, vol. 3, no. 3, pp. 259-66, Jan. 2010.

[28] P. Rocca-Serra et al., "ISA Software Suite: Supporting Standards-Compliant Experimental Annotation and Enabling Curation at the Community Level.," *Bioinformatics (Oxford, England)*, vol. 26, no. 18, pp. 2354-6, Sep. 2010.