

Evaluating (linked) metadata transformations across cultural heritage domains

Kim Tallerås¹, David Massey¹, Anne-Stine Ruud Husevåg¹, Michael Preminger¹, and Nils Pharo¹

Oslo and Akershus University College of Applied Science

Abstract. This paper describes an approach to the evaluation of different aspects in the transformation of existing metadata into Linked data-compliant knowledge bases. At Oslo and Akershus University College of Applied Science, in the TORCH project, we are working on three different experimental case studies on extraction and mapping of broadcasting data and the interlinking of these with transformed library data. The case studies are investigating problems of heterogeneity and ambiguity in and between the domains, as well as problems arising in the interlinking process. The proposed approach makes it possible to collaborate on evaluation across different experiments, and to rationalize and streamline the process.

1 Introduction

Cultural heritage domains have recently experienced substantial efforts in developing new metadata standards intended to increase usability and to enable integration of related resources across established "data silos". In many of the domains, such as in the library community and in broadcasting institutions, these efforts tend to involve Linked data technologies and principles.

The huge amount of existing data produced in compliance with dated standards, requires a significant investigation into transformation processes. In this paper, we describe an approach for the evaluation of different aspects in the transformation of existing metadata into Linked data-compliant knowledge bases. The approach has emerged from work on three partially overlapping and on-going case studies at the Oslo and Akershus University College of Applied Science:

1. The mapping of bibliographic (MARC) records to newly developed ontologies in the library community
2. The (automated) extraction of metadata from semi-structured archive records at the Norwegian Broadcasting Corporation and
3. The interlinking of shared entities across the two domains.

While these case studies have different goals, they share a need for a standardized set of rules for evaluation of performance in a broader context than

traditional evaluation of information retrieval and ontologies represent. The presented approach builds upon existing evaluation principles and metrics, but rationalizes these into a coherent and minimalist system of applicable data sets, representing ground truths for a variety of tasks.

The paper consists of two parts. Firstly, we describe a generic transformation process and provide definitions of the key concepts used in the paper. Secondly, we present the ongoing case studies and the evaluation approach.

2 The road to Linked data - key concepts and processes

Linked data is a set of best practice guidelines for the publishing and interlinking of data on the web, recommending the use of standards such as RDF, URI's and OWL ([1,2]). The publishing and interlinking of legacy data (which is the problem context of this paper) must overcome a variety of heterogeneity conflicts between legacy sources. The conflicts can be structural (caused by disparate modelling approaches) or they can concern inconsistencies in the data (caused by typos, local or changing registration practises, ambiguous name forms, schema flexibility etc.) ([3]). Figure 1 illustrates the process of transforming metadata collections into interlinked knowledge bases. In the figure, "Source schema(s)" denotes any metadata standard or rules for content descriptions. The "Target ontology" can be any formal ontology, providing sets of classes, properties and restrictions. The resulting "Knowledge base" denotes a data set of instances transformed in compliance with the target ontology. According to the Linked data guidelines the target ontology should be based on a formal ontology language such as OWL and the knowledge base should be formalized as RDF.

The transformation process primarily consists of three complementary activities:

- Mapping: Structural transformations based on semantic correspondences between the source schema and the target ontology.
- Extraction: Content transformation consisting of entity and relationship recognition and disambiguation (i.e. information extraction) from textual fields within the metadata.
- Interlinking: The linking of identical entities that are members of different data sets. In Figure 1 $A' \cap B'$ is the intersection of entities that belongs to both (data) set A' and B'. In this context, entities must be understood as "things-that-exist" in the real world. The representation of entities can differ in the data sets, but as long as a unique identifier is provided for each entity in each set, we can formally relate the entities with proper OWL properties.

3 Case studies

At Oslo and Akershus University College of Applied Science (in the TORCH project¹) we are working on three case studies that are focusing on metadata

¹ The TORCH project is an activity of the research group Information systems based on metadata: <http://tinyurl.com/k8gf7dr>

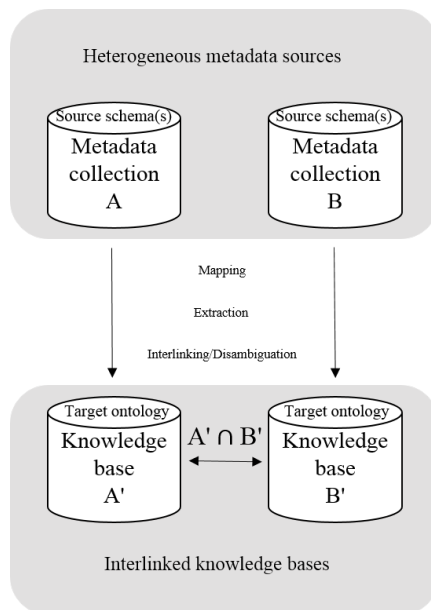


Fig. 1. Overview of a generic transformation process from two sources with related data.

transformations in the library and broadcasting domains. The case studies are investigating problems of heterogeneity and ambiguity in and between the domains and problems arising in the interlinking process.

In the library community a huge amount of the metadata are generated in accordance with established international standards. The two most prominent standards are the suite of MARC metadata schemas and the cataloging code AACR2. MARC has been in use since the late 1960s and was developed as a tool to make the card catalogue machine-readable in order to support metadata exchange between libraries. AACR2 provides rules for the registration of content. Together these standards primarily provide coded fields with string based information intended for human consumption (an inheritance from the card catalogue, see Figure 4 for an example record and [4] for a more detailed discussion of the history and dated features of MARC/AACR2). A "flat" record model (the format was developed prior to both relational database theory and the web) and the string orientation pose severe challenges to the transformation into a data oriented and graph based OWL/RDF environment.

In our case study, we are using a subset of MARC records from the Norwegian National Bibliography. One of the studies is experimenting with mapping based transformations into Knowledge bases compliant with a series of new ontologies

provided in the library community such as BIBFRAME², Schema.org³ and FRBRoo⁴.

The archives of the Norwegian Broadcasting Corporation (NRK) have registered metadata describing TV and radio programs since the early 1990s. While the schema and rules that constrain library MARC data are based on (inter)national standards, NRK's archival metadata is created according to schema and rules developed in-house generating so-called SIFT-records (Searching In FreeText, [5]). A major part of the metadata are free text fields that describe the contents of the programs. Valuable entities, such as people, places and events are hidden within the ambiguity of these natural language descriptions, hampering machine processing, and consequently retrieval and interlinking. One of the cases in the project focuses on the evaluation of methods for extracting these entities and the relationships between them.

Ontologies play a pivotal role in the project. We have been studying existing ontologies within the broadcasting domain, such as the BBC Programmes ontology⁵ and EBU-Core⁶, but felt that they did not fill our needs to describe entities and relationships between entities that we found in the original NRK metadata collection. This was due to our desire to describe both formal elements related to broadcasting (e.g. the relationships between programs, episodes and series) as well as details from the program content (e.g. different kinds of creative works and their creators mentioned or included in a program). Recently Schema.org ([6]) has been extended with elements for the description of broadcasting resources that brought us closer to our required coverage, but we have still felt the need to develop our own broadcasting ontology (TORCH ontology). The ontology is very much inspired by the aforementioned ontologies, and contains mappings to equivalent classes and properties in these. With 50 classes and 60 properties it is not as big as many of the established cultural heritage ontologies, but the design reflects the SIFT-specific needs regarding coverage and supports its two main goals, firstly to be the target of the automated extraction and thus serve as a model for the resulting knowledge base, and secondly to support the manual annotation described below.

Figure 2 illustrates the different case studies and how they interrelate. Dealing with the relatively structured library data, we are primarily concerned with problems related to structured mapping and the outcome of transformations based on such mappings. In the case of broadcasting data, dealing with semi-structured data closer to natural language, we are concerned with problems related to extraction algorithms. In both cases, we aim at disambiguating and interlink the resulting data that are related, based on established tools and experimental algorithms.

² <http://bibframe.org/>

³ <http://schema.org>

⁴ http://www.cidoc-crm.org/frbr_inro.html

⁵ http://www.cidoc-crm.org/frbr_inro.html

⁶ <http://www.ebu.ch/metadata/ontologies/ebucore/>

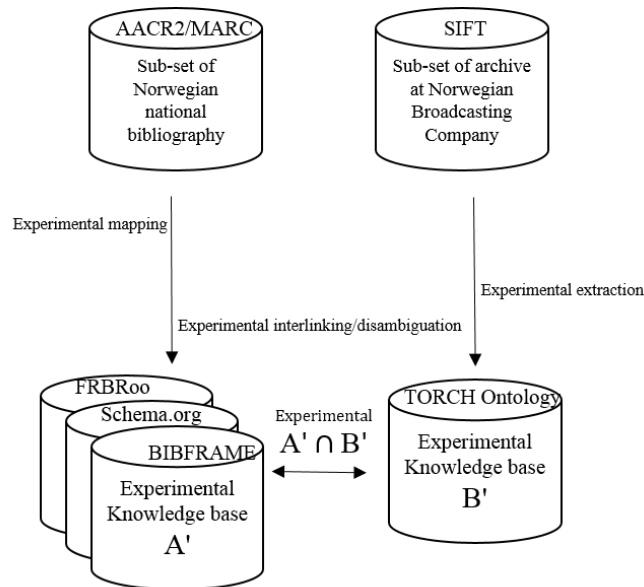


Fig. 2. Overview of the different experimental case studies.

4 Automated extraction and interlinking - ongoing experiments

In the following, we will illustrate and give some examples from the experimental case studies on broadcasting data and the interlinking of these with transformed library data.

A simple prototype for extraction has been developed, that currently consists of a pipeline with three modules: Tokenizing and part-of-speech recognition based on the Oslo-Bergen Tagger⁷; identification of SIFT-specific formatting patterns; and lookup in local gazetteer files, currently Norwegian first and last person names.

The result of the three modules is an array of features for each token. Candidate entities are identified by rules that combine these features and external Linked Open Data and web services are queried to strengthen the evidence.

The identified entities can be used to populate a knowledge base compliant with the TORCH ontology. In order to automatically generate the proper relationships between the entities (e.g between the author and his novel or between the interviewer and her interviewee) internal and external evidence must be collected and analysed. While library data can be used to support the creator-relationship between Eggen and Hilal (see e.g. MARC record in Figure 4), it does

⁷ <http://www.tekstlab.uio.no/obt-ny/english/index.html>

not help to identify the relationship between Eggen and Bratholm that might be dependent on internal (con)textual features.

Figure 3 shows an example of a typical SIFT record and Figure 4 a related MARC record, both with highlighted entities. Figure 5 shows two interlinked RDF graphs based on identified entities and relationships in the SIFT record (the white nodes), and the MARC record (the grey nodes). Corresponding entities in the two graphs are interlinked with the owl:sameAs property. The SIFT records are transformed into a graph compliant with the TORCH ontology and the MARC data is made compliant with the BIBFRAME ontology. Both sets of data were mapped and transformed with the data integration tool Karma [7]. The SIFT data was transformed from the result of the extraction process described above. The MARC data was transformed directly from the record. The project is at a very early stage regarding experimentation on interlinking. The example in Figure 5 is developed manually to illustrate a potential result.

```
Et dypdykk ned i Torgrim Eggens forfatterskap, forfattersjel og
forfatterliv. Inspirert publikum og høy stemning på Rockefeller i Oslo.
(Opptaksdato 960212).
```

```
Programleder Eva BRATHOLM (mv) introduserer kveldens forfatter Torgrim EGGEN
(mv). Intv., innimellom applaus og latter fra salen. Hovedemne er Eggens
siste bok "Hilal", innvandrere, muslimsk kultur, fremmedfrykt, islam ....
Sluttekst.
```

Fig. 3. Excerpts from the content field in a SIFT record. The entities found by the extraction prototype are highlighted. Rockefeller is the venue of the show, located in the city of Oslo. Eva Bratholm is the host interviewing Torgrim Eggen about his novel "Hilal".

```
=100 $aEggen, Torgrim
=24510$aHilal$broman$cTorgrim Eggen
=260 $a[Oslo]$bGyldendal$c2000
```

Fig. 4. Excerpts from a related MARC record from the Norwegian national bibliography. Entities are manually highlighted. "Eggen, Torgrim" is an author, "Hilal" is the title, "Oslo" is the place of publication" and "Gyldendal" the publisher.

5 Evaluation approach

In order to evaluate the experiments described in the previous section, we generate three sets of ground truth data. Figure 6 illustrates how the ground truth

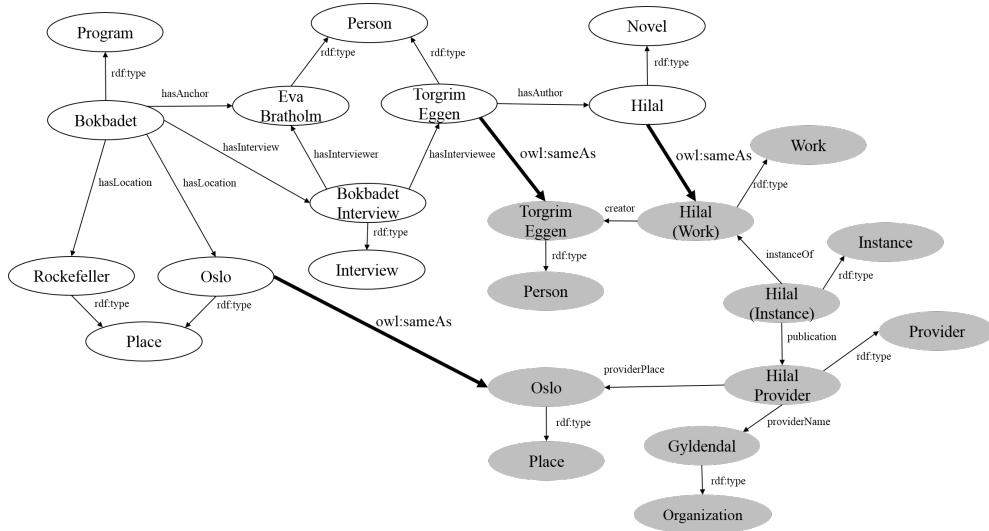


Fig. 5. Two interlinked RDF graphs based on entities in Fig. 3 and 4.

sets, based on selected corpora, are covering the evaluation of each metadata transformation (A and B), respectively, and the eventual interlinking between them. On a conceptual level, these three sets form a coherent approach for the evaluation of metadata transformations and interlinking.

Based on the specific case studies in the TORCH project, we are using this set up for the evaluation of what is described as three complementary activities in Section 2; mapping, extraction and interlinking. The evaluation approach is based on three comparisons between the resulting data from the experimental runs and the ground truth data (our TORCH-specific interests and perspectives are included in parentheses):

- experimental knowledge base A \rightarrow ground truth A (mapping)
- experimental knowledge base B \rightarrow ground truth B (extraction)
- experimental $A \cap B \rightarrow$ ground truth $A \cap B$ (interlinking)

The ground truth data result from (semi-)manual annotations of entities and relationships in the corpus data. In the following, we will briefly describe some of the practical tools we have developed in the TORCH project for the generation of the ground truth data.

5.1 Corpora

To secure a satisfactory level of variety, our corpus of broadcasting data was selected from two different categories of programs; culture and news. 100 SIFT-records from program series in the two categories were chosen. We manually harvested two MARC records related to each of the selected programs from the

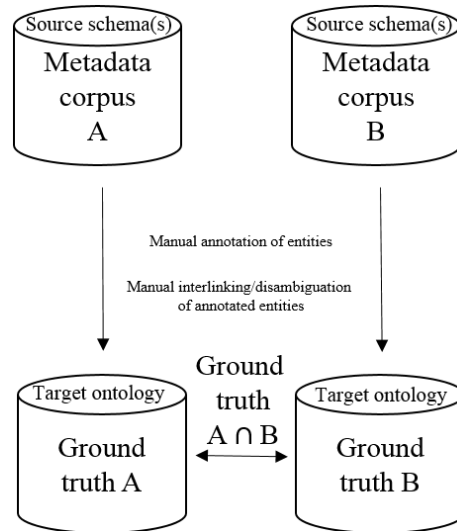









Fig. 6. Illustration of the evaluation approach consisting of three ground truth sets.

Norwegian national bibliography, based on entities found in the SIFT records. This procedure also helped to secure an intersection of entities for the evaluation of interlinking.

5.2 Manual annotation

After reviewing a couple of available annotation tools, we chose to develop our own tool, particularly to gain better control of the different aspects of annotation specific to our project. Annotation productivity was also an important consideration here. The annotation tool consists of a PHP-based GUI supported by relational database structure for both parameterization and persistence. Figure 7 shows a part of the GUI view. Eventually we use Karma to transform annotation data from the relational database into RDF. The GUI allows annotators

- to highlight and classify mentions with classes from the annotation ontology described below. There is also an automatic suggest-and-select feature for linking entities to Wikipedia articles based on the mediaWiki API⁸ search operation.
- to express relations between already classified mentions, using properties from the same ontology(see the table at the bottom of Figure 7). Two special features here are; firstly mentions can relate to the automatically annotated representation of the program (annotated record) itself, and secondly the linking of different mentions of the same entity through a special property, "identicalTo".

 Torgrim EGGEN
 (mv) . Intv ., innimellom applaus og latter fra salen
 Hovedemne er  Eggens siste bok  "Hilal" , innvandrere ,
 muslimsk kultur
 fremmedfrykt , islam
 Reidar LARSEN framf
 "Mr . Understanding and Miss Understood"  (R . Larsen /
 T . Burton) 438"
 og  "Have a little talk with yourself" (R . Larsen / T .
 Burton) 534"
 Sluttekst

Subjekt	Relasjon	Objekt	Slett
Bokbadet_1996/16215:Bokbadet_19960913	hasAnchor	Eva BRATHOLM	<u>x</u>
(R . Larsen	identicalTo	Reidar LARSEN	<u>x</u>

Fig. 7. The annotation tool's main GUI window. Each classified mention in the running text has a special icon to its left, double-clicking on which allows the annotator to establish a relation between it and another, pre-classified mention. The classes, and other information about a mention, are available in a menu window, not shown here, upon selecting a menu or classifying a new mention. The first relation in the table below the running text is between the record (program) representation and the anchor, the second relates two mentions of the same entity to each other.

The annotation tool uses ontologies to provide classes and properties for classification and expression of relationships. In the case of the broadcasting data, we developed the TORCH ontology described in Section 3 partly for this purpose. The classes and the properties in the ontology are used directly to classify entities and relationships between them. In addition to the aforementioned needs concerning coverage, the exposure to test-annotators at an early stage further encouraged the development of a project-specific ontology. The ontology is designed for efficient and consistent annotation by reducing complexity and the intuitive naming of classes and properties [8]. The ontology builds on hierarchies of classes and properties, realized through the RDF Schema properties `rdfs:subPropertyOf` and `rdfs:subClassOf` (e.g. `FictionalCharacter rdfs:subClassOf Person rdfs:subClassOf Agent`). This allows for manual annotations on specific levels, which indirectly and at the same time implies general classifications. Such features can be useful e.g. in order to measure the depth obtained by an extraction algorithm.

⁸ http://www.mediawiki.org/wiki/API:Main_page

High quality manual annotation with a high degree of inter-annotator agreement is dependent on guidelines. With some modifications and extensions, we have based our guidelines on work done by Jonsdottir and the Textlab at University of Oslo [9]. The ontology, guidelines and annotation tool have been developed in an iterative process using a group of LIS students as test annotators and domain experts from the NRK archive as a reference group for ontology development.

In order to generate a ground truth set for the MARC data, following the A path in the figures above, we could use the annotation tool with a bibliographic ontology as input. This would have been especially useful (and interesting) for an analysis of the semi-structured description fields (5XX-fields in MARC parlance). Due to our interest in the results of structured mappings, we have nevertheless chosen to experiment with a straightforward RDF serialization of MARC⁹. The chosen approach secures lossless data and semantics for the comparison with the results of transformations based on other ontologies. There are issues with serializing MARC directly as RDF though, for instance related to the handling of string values, but this part of the project is very much work in progress and the issues will need to be discussed in detail elsewhere. This approach can be considered semi-manual. Data are transformed automatically based on mappings, but quality is assessed and corrected manually afterwards.

5.3 Manual interlinking and disambiguation

The third ground truth set, $A \frown B$, consists of links between corresponding entities in the two sets described above (A and B). The manually created links can be used for the evaluation of (automated) interlinking between RDF graphs, but also to support algorithmic disambiguation of entities as part of the extraction process. This set could be represented in many ways; in our project, we are using the RDF Alignment format¹⁰.

5.4 Evaluation metrics

Evaluation in our context is measuring the correspondence between the result of mappings and automatic extractions on one hand, and the manually developed sets of ground truths on the other hand. In the case studies, we are utilizing established metrics originating from traditional information retrieval and ontology evaluations.

The literature of ontology evaluations is pointing in many directions referencing a variety of metrics concerning everything from design complexity and coverage to usability and human reception. In our context, we are mainly concerned with the level of semantic interoperability between two metadata systems

⁹ The serialization is partly based on the efforts in mapping MARC(21) into RDF found at the metadata registry: <http://marc21rdf.info/>

¹⁰ <http://alignapi.gforge.inria.fr/format.html>

[3], measuring loss and gain of information during transformation from one system to another. In the TORCH project, semantic interoperability is investigated from the perspectives of typical heterogeneity conflicts such as inconsistent string data and structural and semantic variations.

Three metrics (originating from evaluations of information retrieval systems) have dominated information extraction campaigns such as Message Understanding Conferences (MUC, see e.g. [10]); recall, precision and F-score. With the introduction of ontology based information extraction, additional metrics exploiting the features of graphs and ontologies have been used and suggested (see e.g. [11] for an overview). Adaptations of the recall and precision oriented metrics are also common in the evaluation of interlinking (see e.g. the Instance matching track in the ongoing Ontology Evaluation Initiative and [12] for definitions of the metrics in this specific context).

6 Summary and concluding remarks

This paper describes various case studies dependent on systems for evaluation. Such systems exist separately for the evaluation of problems represented by each case study. As our case studies are related through the selection of (corpus) data and the people working with them, we wanted to coordinate such systems in one efficient framework in terms of development and reuse across the studies. We believe that the approach described, consisting of two sets of ground truth data that represent golden standards for the transformations of metadata to RDF, and a third set, consisting of relationships between the previous two, are both efficient and hospitable to (re)use across a variety of problems.

In order to evaluate the experiments, we had to generate three sets of ground truth data. The ground truth data are results of (semi-)manual annotations of entities and relationships in the metadata corpus A and B, and the third ground truth set, $A \cap B$, consists of links between corresponding entities in the two sets A and B. The manually created links can be used for the evaluation of interlinking between RDF graphs, but also to support algorithmic disambiguation of entities as part of the extraction process.

We chose to develop our own annotation tool to annotate entities and relationships in the corpora for the generation of the ground truth data. This allows for manual annotations on specific levels adapted to our projects, which can be useful e.g. in order to evaluate the degree of specificity a certain extraction algorithm is able to achieve. The annotation tool uses ontologies to provide classes and properties for classification and expression of relationships. The ontology and annotation tool have been developed in an iterative process, involving domain experts from the Norwegian Broadcasting (NRK) archive and LIS students as test annotators. The evaluation is based on comparisons between the result of automatic extractions and mappings, and the manually developed sets of ground truth data. This approach can be considered semi-manual. Data are transformed automatically based on mapping, but quality is assessed and corrected manually afterwards. We are mainly concerned about the level of se-

mantic interoperability between two metadata systems, and we are measuring loss and gain of information during transformation from one system to another, using established metrics such as recall, precision and F-score, originating from traditional information retrieval and ontology evaluations.

The case studies described above are still work in progress and will be documented in detail elsewhere. Regarding the further development of the approach, we are looking into the potential provided by the ontology-based features. Using ontologies as the basis of ground truth data gives us the opportunity to evaluate both transformed entities and the relationships between them. In practice, this implies the opportunity to move from entity recognition to information extraction without adding further annotations. We have also mentioned the possibility of exploiting the hierarchical structures in the ontology, for evaluating the specificity-ability of extraction algorithms. As future work we wish to investigate how ontologies can be further exploited as a basis for evaluating metadata transformations.

References

1. Berners-Lee, T.: Linked data: design issues. Technical report, W3C (2006)
2. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a global data space. Morgan & Claypool (2011)
3. Tallersås, K.: From many records to one graph: Heterogeneity conflicts in the Linked data restructuring cycle. *Information Research* **18**(3) (2013)
4. Svenonius, E.: The Intellectual Foundation of Information Organization. The MIT Press, Cambridge, Massachusetts (2000)
5. Reigem, O.: Sift - searching in free text: A text retrieval system (abstract only). *SIGSOC Bull.* **12-13**(4-1) (May 1981) 59–
6. Raimond, Y.: Schema.org for TV and Radio markup (2013)
7. Knoblock, C., Szekely, P., Ambite, J., Gupta, S., Goel, A., Muslea, M., Lerman, K., Taheriyan, M., Mallick, P.: Semi-automatically mapping structured sources into the semantic web. In: *Proceedings of the Extended Semantic Web Conference, Crete, Greece (2012)*
8. Hinze, A., Heese, R., Luczak-Rösch, M., Paschke, A.: Semantic Enrichment by Non-experts: Usability of Manual Annotation Tools. In: *Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E., eds.: The Semantic Web ISWC 2012. Volume 7649 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 165–181*
9. Jonsdottir, A.: ARNER, what kind of name is that?: an automatic rule-based named entity recognizer for Norwegian. PhD thesis, University of Oslo (2003)
10. Grishman, R., Sundheim, B.: Design of the MUC-6 evaluation. In: *Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996, Association for Computational Linguistics (1996) 413–422*
11. Maynard, D., Peters, W., Li, Y.: Metrics for Evaluation of Ontology-based Information Extraction. In: *WWW Conference 2006, Workshop on "Evaluation of Ontologies for the Web", Edinburgh, Scotland (2006)*
12. Ferrara, A., Lorusso, D., Montanelli, S., Varese, G.: Towards a benchmark for instance matching. In: *Proceedings of the 3rd International Workshop on Ontology Matching (OM-2008). (2008)*