Consuming Linked Closed Data

Marcus Cobden, Jennifer Black, Nicholas Gibbins, and Nigel Shadbolt

{mc08r,jlb08r,nmg,nrs}@ecs.soton.ac.uk
School of Electronics and Computer Science,
University of Southampton, UK

Abstract. The growth of the Linked Data corpus will eventually prevent all but the most determined of consumers from including every Linked Dataset in a single undertaking. In addition, we anticipate that the need for effective revenue models for Linked Data publishing will spur the rise of Linked Closed Data, where access to datasets is restricted. We argue that these impeding changes necessitate an overhaul of our current practices for consuming Linked Data. To this end, we propose a model for consuming Linked Data, built on the notion of continuous Information Quality assessment, which brings together a range of existing research and highlights a number of avenues for future work.

1 Introduction

Research among the Semantic Web's Linked Data community has thus far focused on techniques and strategies for publishing and consuming Linked Open Data (LOD): data which is available free of charge and without access restrictions. As Semantic Web technologies see further adoption, more linked Resource Description Framework (RDF) datasets are published, and companies seek effective revenue models for Linked Data publishing, it is increasingly likely that datasets, to which access is restricted, will be published. It is, therefore, of increasing importance that research in this area considers how to consume Linked Closed Data (LCD) as well as its counterpart Linked Open Data.

Conventional online content publishers generally rely on advertising revenue to generate a return on their investment in content; however, it is not clear whether an advertising-based revenue model will be viable for RDF data. Adverts incorporated into raw data will either be easily identifiable among the data items, and thus easily removable, or indistinguishable, and may thereby undermine the integrity of the data[1]. Further research is required in order to conclusively demonstrate or refute the viability of advertising supported Semantic Web data publishing.

Very few commercial content providers are able to publish content for free, as it remains costly to produce; that it is cheap to reproduce content in the digital age is immaterial. Unless an organisation has the benefit of donated labour (e.g. Wikipedia[2]), receives subsidies or is able to write off the investment in some way

[1] http://www.ldodds.com/blog/2010/01/thoughts-on-linked-data-business-models/
[2] http://www.wikipedia.org/

(such as goodwill or marketing), a suitable revenue model for publishing Linked Data will be sought.

It has been speculated that World Wide Web content will move away from ad-supported revenue models towards paid-access revenue models [20]: recently The Times newspaper, along with others such as the New York Times, has taken the step of gating access to its online content to paid subscribers, rather than supporting its online content with advertising [7,28]. If these moves signal the beginning of a trend away from advertising-supported content, it is possible that commercial providers of Linked Data will also wish to explore other revenue generation possibilities.

The loss-leader approach is an alternative to advertising supported models, it aims to use free offerings to attract customers to purchase other products and services which an organisation offers. The Freemium revenue model generates revenue from free content by charging for a premium counterpart, relying on use of the free content to drive sales of the premium version [2]. Freemium may be considered a specialised form of loss leader, where the free and premium products are more strictly connected. This is less risky than a simple loss leader model as it is closer to a pure paid access model, and the scope of the free offering or the price of the premium version can be adjusted to compensate for the impact of freeloading and to respond to other market forces. If the Freemium model is not viable, then it is only a small change to revert to a paid access model.

In economic terms, LOD is a 'public good' [29]: a good which is non-rivalrous (consumption of one good does not reduce the availability of the good for others) and non-excludable (no-one can be excluded from consuming the good) [21]. Public goods, being non excludable in nature, are difficult to charge for, as there is no means to prevent free-riding [12] — the act of consuming more than one's fair share, or shouldering less than the fair share of the production costs (only the latter applies in the case of non-rival public goods).

In contrast to public goods, 'club goods' are those which remain non-rivalrous but *are* excludable [6,21]. The premium content of the Freemium model may be considered a club good, while the free version remains a public good. The viability of the Freemium model depends on the degree of free-riding and the level of premium sales promoted by free content.

Regardless of this, LCD will arise from the need to publish Linked Datasets which remain excludable goods. Ultimately, if advertising-supported Linked Data publishing is not a success, the commercial viability of publishing Linked Data hinges on whether consumers of Linked Data will be willing to pay for access to datasets. There is evidence to suggest that people remain unwilling to pay for online content and are unlikely to change [5], although it has been argued that this unwillingness to pay is a result of the widespread availability of free content online [9].

In addition to the challenge presented by closed datasets, the growth of the Semantic Web and the increasing number of available datasets presents further challenges. Existing Semantic Web storage, query processing and reasoning technologies do not exhibit sufficient scalability to match Web-scale growth [11].

Therefore, it will at some point become unfeasible for an agent to use all datasets relevant to a particular task, and as a result, Semantic Web agents will need to be able to select a subset of datasets from the Web of Data when planning how to complete a particular task. Aggregation datasets may alleviate this to some degree, but specialised niche or premium datasets will still require consideration on an individual basis. Thus, if we are to fully harness the potential of Linked Data we must construct a coherent model for consuming both Open and Closed data.

Given this, in this paper, we describe a motivating scenario for our work, which serves to highlight some of the issues we expect to see in the future (Section 2). We then explore the challenges we identified in the introductory section in greater detail (Section 3), and go on to propose a model for the consumption of both Open and Closed Linked Data (Section 4). Finally, we conclude with a summary of the avenues for future work which we have identified (Section 5).

2 Motivating Scenario

In order to better explain the issues LCD presents, we consider the following scenario.

Having just concluded her undergraduate degree in Computer Science, Lucy has decided to pursue further study and is looking into the options offered by different universities. She is aware of the Masters courses offered by her current institution but knows nothing about those of others. Searching online she finds a discussion forum on the subject, where members have posted their experiences and opinions of different courses, a list of rankings compiled by a national newspaper, and the results of a national student and graduate survey.

Her first choice would be the newspaper league tables, as she has heard favourable words about previous revisions, however she is put off by the access charge so decides to first explore the free options. Lucy skims the discussion forum but is struck by the overall predominance of poor or unfavourable reviews. She is suspicious of this bias, suspecting that there are factors at work that she is unaware of, and decides that she cannot rely on the discussion forum as a source of reliable information. Lucy finds the survey very informative, it provides good indications on the level of facilities of each university, how pleased the students were with the teaching, and a good estimate of the worth of a qualification from each institution in the earnings of its graduates.

However, Lucy is considering is studying for a doctorate following her Masters degree, and as a result she is also interested in universities scoring highly for research in her field. Lucy decides to pay for access to the newspaper league table as the other sources did not have information on research rankings, and it comes from a highly reputable source, so Lucy feels confident in trusting it.

In our scenario, the agent, in this case Lucy, is tasked with learning about Masters level degree programmes open to Computer Science graduates. The scenario has strong parallels with the world of Linked Data — a practically identical scenario could be described involving a Semantic Web agent and a number of Linked Data datasets. The agent finds a number of data sources, of

varying calibre and cost, and chooses between them. As time is not a critical factor, our agent can afford to explore the free options and decide later whether or not it needs to purchase the premium datasets. In this instance, the first dataset was of low quality, and the second was good quality but did not contain sufficient information.

3 Challenges for LOD and LCD

Building on this scenario, we have identified a number of important challenges which we believe will become increasingly prominent in the future of the Web of Data. They arise from the need for automated methods for consuming Linked Data, considerations which we expect LCD to require and the growth we may see in the number of published Linked Datasets.

3.1 Dataset Discovery

Dataset discovery, the problem of finding previously unknown datasets on the Web, is perhaps the simplest of the aforementioned challenges. Automatic dataset discovery will be needed if we wish applications to find serendipitous information completely autonomously, as manual identification of relevant datasets will only scale so far.

As the Web is an unmoderated and decentralised network, there will never be a definitive, centralised solution to this problem. However, simple mechanisms can go a long way towards providing an effective solution. URI resolution and inter-dataset link traversal is perhaps the most common automatic method currently in use: it may be considered the Linked Data equivalent of Web crawling, which follows Hypertext links rather than RDF links. Resolution and traversal has the same pitfalls as Web crawling, relying on incoming links in order to discover new areas, and so, it performs poorly at finding datasets which are new to the corpus and have zero, or very few, incoming links.

While inter-dataset link traversal will only find datasets which have incoming links from the known corpus, a hybrid approach, crawling both the Hypertext Markup Language (HTML) Web and the data Web, may be able to locate new datasets via incoming links from HTML pages. This may have more success locating new datasets as it is, arguably, currently easier to share new links on the Web than to arrange for links to be added to other datasets. Unfortunately, Web crawling is a non-trivial task, given estimates of the size and growth rate of the Web [16,4].

Figure 1 illustrates the discoverability of different datasets through link traversal: datasets b and d are not discoverable through RDF link traversal as they receive no incoming links from the known Web of data. Both b and d are potentially discoverable through HTML link traversal, as they are linked to from HTML documents a and c respectively. However, only dataset d is actually discoverable through HTML link traversal, as the HTML document c is linked to from other documents, whereas a is not.

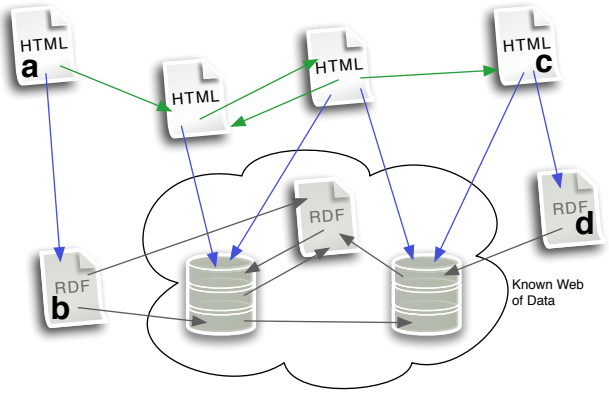

Fig. 1. Illustrating dataset discovery through linkae

While Web-crawling may not be a realistic option for most, crawled indexes remain a good source for dataset discovery. Semantic Web search engines (such as Swoogle[3] [8] or Sindice[4] [23]) are, therefore, a useful tool for dataset discovery, although whether they contain the latest datasets will depend entirely on the resources of the search engine.

The Vocabulary of Interlinked Datasets (voiD) [1], Semantic Sitemaps, and conventional search engine sitemaps enable the description of the contents of sites and datasets. Apart from voiD, such sitemaps are not directly of use for dataset discovery (they are commonly manually submitted to search engines): voiD allows one to state that a dataset contains a given number of links to another dataset.

Dataset directories offer another avenue for dataset discovery: a small number of Community-maintained dataset lists exist[5], however, they lack a uniform formal markup and the levels of metadata vary.

3.2 Information Quality Assessment

In the Web of open data, the challenge of Information Quality (IQ) Assessment is critical: information sources may vary substantially in every aspect, and the key to obtaining useful information is the ability to distinguish high quality information from low [22]. The same is true of the Web of HTML documents, as we described in Section 2. Simply adopting the perceived trustworthiness of its source as the sole indicator of the quality of some information fails to consider the full complexities of the Web of data, where information may have multiple sources, may have been re-published, or may be derived from other information

[3] `http://swoogle.umbc.edu/`
[4] `http://www.sindice.com/`
[5] `http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/DataSets`, `http://ckan.net/tag/lod`

[17]. Information quality will be especially important for LCD, as the costs of access further motivate the desire for high-quality information. Other Semantic Web research frontiers, such as the e-Science movement, also have cause to value IQ highly [24].

IQ is an inherently subjective measure: an IQ assessment may be dependent on both the agent performing the assessment, and the intended use of the data. However, the information on which individual IQ criteria are assessed is not generally subjective, and so it is worthwhile to publish such information in order to enable others to make better decisions. Existing dataset metadata vocabularies offer little in the way of IQ information — for example: voiD and the Statistical Core Vocabulary (SCOVO) [18] allow one to declare the number of links between two datasets, however, they fail to describe the distribution of the links within the dataset. There is a significant difference between a dataset where all outgoing links are from a small collection of the whole, and another in which they are evenly distributed.

Hartig proposes to classify the influences of belief decisions into three categories: i) information quality, ii) provenance, and iii) others' opinions [17]. We argue that, instead, IQ should be the sole determinant of belief. Provenance and peer opinions should be incorporated into IQ criteria as they require the same manner of decision as regular IQ criteria. Different applications presumably have different provenance rigorousness requirements, and the valuation of different agents' opinions will be unique to the agent and the task context. We do not wish to dismiss the importance of provenance; rather we contend that it makes sense to assess it alongside other IQ criteria.

3.3 Dataset Selection

Inadequate scalability on the part of storage, query processing and reasoning technologies will necessitate the rise of limited-rationality systems, which are designed to make decisions without complete information [11]. Aggregation services and techniques may help Linked Data consumers cope with the growing number of sources but, ultimately, surveying a significant proportion of datasets in a single undertaking will soon cease to be feasible for most, if not all, applications. Therefore, the challenge of dataset selection — selecting datasets from a larger collection of potentially relevant datasets — will become even more important. Note that we consider the curation of this larger collection of datasets to be a separate challenge, one of dataset discovery, which we describe in Section 3.1.

Selection methods will also have to address the challenge posed by gated datasets: if access to a dataset cannot be assumed, then selection criteria will have to function without direct experience of the dataset. Metadata on the qualities of each dataset will, then, be crucial to making good decisions, and we can expect reviews of the datasets and their metadata to play an important role in ad-hoc quality control.

3.4 Information Integration

In order to take full advantage of data available across different datasets, it is necessary to integrate them in some way. This problem is not unique to either

open or closed Linked Data; it is an intrinsic challenge of data management. Information integration is also by no means a new challenge: its importance, and complexity, became apparent soon after the rise of enterprise database use [3,19]. Traditional approaches commonly employ a series of mappings in order to translate from the query at hand to an internal representation, and then to a dataset-specific form [13]. Semantic Web technologies offer some improvement on traditional database systems, in that Semantic Web systems share a common data model and schema re-use is easier. However, this does not significantly reduce the heterogeneity of information systems, but instead raises the problem to a higher level, to one of schema integration, or ontology alignment [10].

In addition to schema integration, we also encounter the challenge of co-reference [15]: schema integration manages the problem of class, property and relation equivalence, whereas co-reference is concerned with instance equivalence. For example, two different datasets might both use the same vocabulary to express their data, but this does not stop them from coining different identifiers to denote the same entity. Co-reference services such as sameAs.org[6] [14] maintain and publish co-reference data.

4 Proposed Model

As we have argued above, the growth of the Web of data and the anticipation of gated Linked Data sources necessitates a new approach to consuming Linked Data. Resource-based restrictions and the increasing number of datasets published online will prevent agents from considering every dataset available, forcing them to pick and choose the most effective datasets for their query. Any new approach must be able to make this decision before any datasets are accessed, as doing so may incur costs — monetary, computational or otherwise. There are many criteria on which this decision may be made: noteworthy examples are relevance, quality, cost, and licensing.

We present below an iterative model for the general behaviour of a data selection system built around continuous IQ assessment.

Figure 2 illustrates our model: the solid arrows indicate the normal flow of control between stages, whereas the dashed arrows indicate where control may return to previous stages after some IQ assessment. We explain each stage in turn in the following sections.

4.1 Task Analysis

In this first stage, the task at hand is analysed to gain an understanding of what it requires and, where possible, the context in which it is assigned. Good task analysis is important to ensure the rest of the model operates effectively; in order for an agent to be able to select datasets effectively, it must have a good understanding of the task it has been set. This stage maps the assigned task to an internal representation which captures the requirements and enables subsequent mapping into different data vocabularies.

[6] `http://sameAs.org/`

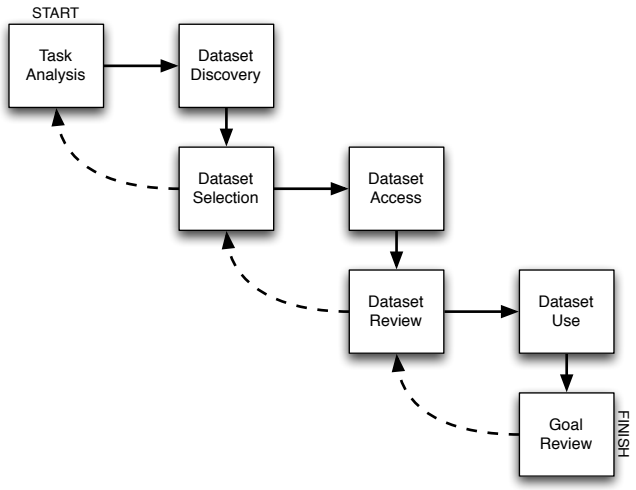

Fig. 2. Flow of the Linked Data consumption process

4.2 Dataset Discovery

In this second stage of the model, the focus is to discover datasets relevant to the task at hand, and to collect metadata about them. An agent may do this in a number of ways, perhaps through Web crawling, querying search engines and dataset directories, from previous experience or by asking its peers. Metadata on each dataset is important for the next part of the process, Dataset Selection, in order to reach a good decision on the subset of the available datasets.

As we discussed in Section 3.1, there is a need for techniques to improve dataset discovery methods. Link traversal performs poorly at finding new datasets, however the addition of 'backlink' notification services can improve this by helping to create incoming links to new datasets, where only outgoing links previously existed [26]. Dataset directories will be of use to discovery, however human maintenance of them may not scale with the growth of the Semantic Web as Web-era content directories[7] have difficulty keeping abreast of the growth of the HTML Web. It is likely that in the future we will have to employ automated methods in order to effectively curate large directories of datasets.

Current dataset-level metadata is limited; the voiD [1] focuses on the description of the links between datasets and has made only a marginal consideration of the description of a dataset's contents. The voiD documentation[8] states that datasets may be categorised by using the `dcterms:subject` property to describe the contents of this dataset. As the subject of the dataset is defined as the union of all `dcterms:subject` assertions, complex subject descriptions must use a more specialised vocabulary, such as Simple Knowledge Organization

[7] `http://www.dmoz.org/`
[8] `http://vocab.deri.ie/void/guide`

System (SKOS). Other dataset-level metadata Web standards such as Semantic Sitemaps or Search engine Sitemaps facilitate some description of datasets, but are focused towards search engine crawling. There is also a case for community submitted metadata, which would allow claims of certain dataset qualities to be corroborated or contested, and poor categorisations improved.

4.3 Dataset Selection

Equipped with metadata on potentially useful datasets, the Dataset Selection phase is to select from among relevant datasets the best possible subset for the task at hand. This stage is also responsible for assessing each dataset in terms of their suitability to the completion of the task.

As we outlined in Section 3.3, agents will need to select between the datasets, weighing the costs and benefits of each. In order for the cost/benefit decision to have a worthwhile result, each dataset must first be evaluated with respect to relevant IQ criteria. We argue that evaluation of datasets must be performed before the datasets are accessed, as doing so may incur time or resource costs, and in the case of gated datasets, financial costs. That is not to say that the consideration cannot take into account existing knowledge of the datasets, rather, that gaining such knowledge is part of a later phase. To do otherwise, accessing and surveying each dataset before evaluation (assuming one cannot afford to buy access to each potentially useful gated dataset), biases the selection in the favour of open data sources (or gated datasets which have been previously accessed).

While the cost of gated datasets may understandably bias selection in favour of free sources, it does not follow that gated sources do not hold valuable or worthwhile data, thus it would be regrettable to exclude them. The cost/benefit decision will need to weigh the cost and potential value of gated datasets against the operational costs and potential values of first exploring free datasets.

Without access to datasets, IQ assessment methods cannot evaluate their value directly, instead the assessments must rely on the metadata gathered on each dataset. This may include reputation-style metadata, describing others' experiences as a means of verifying that the contents of the dataset match the reported qualities. In addition to per-dataset reputation, publishers might also gain a reputation for the quality of datasets which they publish. Understandably, if the agent has first-hand experience of a dataset, it should be free to employ that knowledge in its evaluation.

We anticipate that publishers of gated datasets will employ novel methods in an attempt to offset the bias free datasets enjoy. Techniques such as free trial access, or demonstration datasets, might be employed to build a reputation for the dataset or publisher.

At this stage, the control may return to the Task analysis phase (i) if too few datasets were discovered, or (ii) if the discovered datasets are poor choices. Conceivably, when faced with a poor selection of datasets, an advanced agent might opt to select a dataset which it believes will improve its understanding of the task, rather than one directly applicable to the completion of the task. Having selected the datasets which will be used, the process will continue to the Dataset access phase.

4.4 Dataset Access

During this phase, the executing agent gains access to the previously selected datasets: in the case of gated datasets this is likely to entail some form of authentication and potentially also payment. There are a number of current proposals for Semantic Web authentication protocols, FOAF+SSL [27] and OpenID [25], however, no proposals on the subject of payment, besides the reserved `HTTP 402 Payment Required` status code. In our example (Section 2), the authentication and payment is a function of the newspaper's website. The type of payment model may also vary; possible models include subscription based access and pay-as-you-go micro-payments.

In the case of open datasets, this stage still provides an opportunity to verify technical IQ measures such as availability and latency.

4.5 Dataset Review

The Dataset Review phase brings further IQ assessment. At this point, the agent reviews the datasets which it has selected and gained access to. If it then deems the selection of datasets of insufficient worth for the task, it may return execution to the dataset selection phase.

The IQ assessment of each dataset may be an active task: the agent may actively explore and sample the datasets in order to verify their reported qualities. If the agent is satisfied that the datasets are still of worth, then execution continues in the Data Use phase — if it deems the selection of datasets insufficient, execution may return to the dataset selection phase.

4.6 Data Use

In this phase, the executing agent queries the selected datasets in order to complete the assigned task. The requirements and terminology of the task are mapped from the agents' internal representation to the domain vocabulary of each dataset. The exact execution of this phase will differ with the type of task which the agent has been set. When it is judged that querying of the datasets is complete, the agent enters the goal review phase.

4.7 Goal Review

Finally, the agent evaluates whether the goals of the task have been met. This phase may also include an IQ assessment of the outcome, to ensure sufficient credibility and likelihood. If the criteria of the task are met, the task can be considered complete; if not, like Fensel's simple algorithm [11], we seek further information, returning to earlier stages to do so.

5 Conclusion and Future Work

To summarise, we have argued that the rise of Linked Closed Data is increasingly likely, especially if advertising based revenue models prove to be unviable.

Considering the implications of consuming closed datasets highlights a number of research challenges: i) dataset discovery, ii) information quality assessment, iii) dataset selection and iv) information integration. In more detail, dataset discovery entails the automatic discovery of relevant datasets in the Web of Data, information quality assessment is the evaluation of the quality of information with respect to a certain task, dataset selection involves the nomination of a subset of discovered datasets, for use towards a particular task, and, finally, information integration is the combining of information from different sources and in heterogeneous formats.

In light of this, we present a model for consuming Linked Data which takes into account the challenges above. Our model is built around the notion of continuous Information Quality assessment: we repeatedly evaluate the quality of our sources and their fitness towards the task at hand, so that we can reach a high quality, trusted outcome.

Existing research has already laid the groundwork for our model of consuming Linked Data, but significant research challenges remain: from the overarching challenges of agent-based task planning, to the enabling infrastructure challenges which underlie this model of consuming Linked Data. These infrastructure challenges, such as provenance, reputation and dataset metadata, are those on which we plan to focus our future research. Finally, we also plan to further investigate the willingness of consumers to pay for Linked Data.

References

1. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing Linked Datasets - On the Design and Usage of voiD, the 'Vocabulary of Interlinked Datasets'. In: WWW 2009 Workshop: Linked Data on the Web (LDOW). Madrid, Spain (2009)
2. Anderson, C.: Free: The Future of a Radical Price: The Economics of Abundance and Why Zero Pricing Is Changing the Face of Business. Random House Books (Aug 2009)
3. Batini, C., Lenzerini, M., Navathe, S.B.: A Comparative Analysis of Methodologies for Database Schema Integration. ACM Comput. Surv. 18(4), 323–364 (1986)
4. Bergman, M.K.: The Deep Web: Surfacing Hidden Value. The Journal of Electronic Publishing 7(1) (August 2001)
5. Chyi, H.: Willingness to pay for online news: An empirical study on the viability of the subscription model. Journal of Media Economics 18(2), 131–142 (2005)
6. Cornes, R., Sandler, T.: The Theory of Externalities, Public goods, and Club goods. Cambridge University Press (1996)
7. Dewan, R.M., Freimer, M.L., Zhang, J.: Management and Valuation of Advertisement-Supported Web Sites. Journal of Management Information Systems 19(3), 87–98 (2002)
8. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, S.R., Peng, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: A Search and Metadata Engine for the Semantic Web. In: CIKM '04: Proc. of the thirteenth ACM conf. on Information and knowledge management. pp. 652–659. ACM Press, New York, NY, USA (2004)
9. Dou, W.: Will Internet Users Pay for Online Content? Journal of Advertising Research 44(04), 349–359 (2004)
10. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Berlin (2007)

11. Fensel, D., van Harmelen, F.: Unifying Reasoning and Search to Web Scale. IEEE Internet Computing 11(2), 96–95 (2007)
12. Gaustad, T.: The Problem of Excludability for Media and Entertainment Products in New Electronic Market Channels. Electronic Markets 12(4), 248–251 (2002)
13. Genesereth, M.R., Keller, A.M., Duschka, O.M.: Infomaster: An Information Integration System. In: Proceedings of the 1997 ACM SIGMOD international conf. on Management of data. pp. 539–542. ACM, New York, NY, USA (1997)
14. Glaser, H., Millard, I.: RKBPlatform: Opening up Services in the Web of Data. In: Poster Paper at International Semantic Web Conference 2009 (October 2009)
15. Glaser, H., Millard, I., Jaffri, A., Lewy, T., Dowling, B.: On coreference and the semantic web. In: 7th International Semantic Web Conference (May 2008)
16. Gulli, A., Signorini, A.: The Indexable Web is More than 11.5 Billion Pages. In: WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web. pp. 902–903. ACM, New York, NY, USA (2005)
17. Hartig, O.: Towards a Data-Centric Notion of Trust in the Semantic Web. In: Proceedings of the Second Workshop on Trust and Privacy on the Social and Semantic Web. vol. 576. CEUR-WS.org, Heraklion, Greece (May 2010)
18. Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L., Ayers, D.: SCOVO: Using Statistics on the Web of Data. In: ESWC 2009. vol. 5554, pp. 708–722 (2009)
19. Lenzerini, M.: Data Integration: A Theoretical Perspective. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. pp. 233–246. ACM, New York, NY, USA (2002)
20. Lopes, A., Galletta, D.: Consumer Perceptions and Willingness to Pay for Intrinsically Motivated Online Content. Journal of Management Information Systems 23(2), 203–231 (2006)
21. McNutt, P.: Public Goods and Club Goods. Encyclopedia of Law and Economics (1999)
22. Naumann, F.: Quality-Driven Query Answering for Integrated Information Systems, LNCS, vol. 2261. Springer (2002)
23. Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., Tummarello, G.: Sindice.com: A Document-oriented Lookup Index for Open Linked Data. International Journal of Metadata, Semantics and Ontologies 3(1) (2008)
24. Preece, A.D., Jin, B., Pignotti, E., Missier, P., Embury, S.M., Stead, D., Brown, A.: Managing Information Quality in e-Science Using Semantic Web Technology. In: Sure, Y., Domingue, J. (eds.) ESWC. LNCS, vol. 4011, pp. 472–486. Springer (2006)
25. Recordon, D., Reed, D.: OpenID 2.0: a platform for user-centric identity management. In: DIM '06: Proceedings of the 2nd ACM workshop on Digital identity management. pp. 11–16. ACM, New York, NY, USA (2006)
26. Salvadores, M., Correndo, G., Szomszor, M., Yang, Y., Gibbins, N., Millard, I., Glaser, H., Shadbolt, N.: Domain-Specific Backlinking Services in the Web of Data. Web Intelligence (September 2010)
27. Story, H., Harbulot, B., Jacobi, I., Jones, M.: FOAF+ SSL: RESTful Authentication for the Social Web. In: Proceedings of the First Workshop on Trust and Privacy on the Social and Semantic Web (SPOT2009) (2009)
28. Thurman, N., Herbert, J.: Paid content strategies for news websites: An empirical study of British newspapers' online business models. Journalism 1(2), 208–226 (2007)
29. Varian, H.R.: Markets for information goods. In: Monetary Policy in a World of Knowledge Based Growth: Quality Change and Uncertain Measurement. Palgrave Macmillan (2001)