Part IV

Conclusion

Chapter 14

Conclusion

As the interconnectivity of information systems increases and more and more information becomes available on the Web, the topic of selecting high quality information from the vast amount of accessible information is gaining importance.

This thesis analyzed the concept of information quality and gave an overview of different metrics that can be used to assess information quality in the context of web-based information systems. Information quality is a multidimensional construct. The relevance of the different quality dimensions depends on the task at hand of the information consumer and her subjective preferences. Therefore, this thesis propagates a subjective, user-centric approach to quality-driven information filtering. Instead of having the designer of an information system decide for the user on a single, fixed method to ensure information quality, the user is empowered to express subjective, task-specific filtering policies and, by applying these policies, to decide for herself whether to accept or reject information.

This user-centric approach is explored by developing the WIQA - Information Quality Assessment Framework. The WIQA framework is a set of software components that can be used by web-based information systems in order to enable information consumers to filter information using a wide range of different filtering policies. The WIQA framework has been integrated into a general-purpose web browser and it has been shown how the framework can be applied within a financial information integration scenario.

The main features of the WIQA framework are:

Flexible Representation of Quality-Related Meta-Information.
The framework employs a flexible data model for representing information from the Web together with quality-related meta-information. The data model provides for the integration of heterogeneous information

from multiple sources and allows meta-information to be represented on the level of detail that is required by a specific application.

Support for Quality-Based Information Filtering Policies. The framework enables information consumers to filter information using a wide range of different filtering policies. Policies are expressed using a declarative policy language and may combine different content-, context-, and rating-based information quality assessment metrics. The WIQA framework can be extended with domain-specific assessment metrics as different application domains require different metrics.

Explanation of Filtering Decisions. The key factor for a user to trust the quality of positively filtered information is her understanding of the filtering process. In order to facilitate this understanding, the WIQA framework can generate explanations why information satisfies a given policy.

The following section summarizes the contributions of this thesis and discusses the impact of the work. Afterwards, directions for future research are outlined.

14.1 Contributions

This thesis proposes an innovative solution to quality-driven information filtering in the context of web-based information systems. The main contribution of the thesis is the development of the WIQA - Information Quality Assessment Framework. The development involves the design of the following artifacts:

The Named Graphs Data Model. The RDF data model was extended to the Named Graphs data model in order to ease the representation and the exchange of meta-information about RDF data. Named Graphs provide a high-value but small and incremental change to the RDF recommendations. Combined with further specific vocabulary, Named Graphs will be beneficial in a wide range of application areas and will allow the usage of a common software infrastructure spanning these areas. The Named Graphs data model has been widely discussed in the RDF community. Indicators for the utility of the model are its adoption by the W3C Data Access Working Group (DAWG) as part of the data model behind the SPARQL query language [PS05] and the increasing number of RDF toolkits that implement Named Graphs [BW06].

Syntaxes for Serializing Graph Sets. Chapter 5.3 introduced the TriX and TriG syntaxes for serializing sets of named graphs. Besides the ability to serialize multiple graphs into a single document, the TriX syntax also addresses several drawbacks of the RDF/XML syntax. The TriX syntax corresponds closely to the triple structure of an RDF graph. Therefore, it is less complicated to parse than the RDF/XML syntax and works better with generic XML tools such as Xpath, XSLT, and XQuery.

Semantic Web Publishing Vocabulary. Chapter 6 developed the Semantic Web Publishing Vocabulary (SWP), which provides terms for expressing different degrees of commitment towards published information and for representing digital signatures. Linking information to authorities and optionally assuring these links with digital signatures gives information consumers a secure basis for using filtering policies which rely on information provenance. Signing RDF graphs requires specific canonicalization and digest algorithms. The SWP vocabulary is the first RDF vocabulary that provides terms to identify these algorithms and to describe the combination of algorithms that is used to calculate a signature. This enables the SWP vocabulary to represent serialization-independent signatures and makes it possible to verify signatures even after information from different sources is combined and serialized using a different serialization syntax.

WIQA-PL Policy Language. Chapters 9 and 10 developed the WIQA-PL policy language. WIQA-PL is a declarative language for expressing quality-based information filtering policies. Policies can combine different content-, context-, and rating-based assessment metrics. Policies are expressed in the form of graph patterns and filter clauses. As information quality assessment often requires domain-specific assessment metrics, WIQA-PL provides an extension mechanism that enables domain-specific metrics to be included into policies. WIQA-PL policies may contain explanation templates which are used by the WIQA - Filtering and Explanation Engine to generate natural language as well as RDF explanations about filtering decisions.

Open Source Implementations. The Named Graphs data model, the TriG and TriX syntaxes and the Semantic Web Publishing Vocabulary have been implemented as part of NG4J - Named Graph API for Jena. The WIQA-PL policy language has been implemented as part of the WIQA - Filtering and Explanation Engine. NG4J and the WIQA

engine are both available under open source licenses. NG4J is already being used by several other projects [SS05, WN06].

WIQA Browser. The WIQA browser demonstrates how information quality filtering capabilities can be integrated into a standard web browser. The WIQA browser enables users to extract structured information from web pages. Extracted information can be filtered using WIQA-PL policies. When a policy is applied, an 'Oh, yeah?'-button appears next to each piece of information that is displayed by the browser. Pressing this button opens a new window with an explanation why the piece of information fulfills the selected policy. I have discussed the WIQA browser with Tim-Berners Lee, the inventor of the World Wide Web, and Daniel Weitzner, the director of the W3C Technology and Society activities. Tim-Berners Lee confirmed that the WIQA browser is the first implementation of his idea of the 'Oh, yeah?'-button he knows about. Daniel Weitzner presented the WIQA browser in his keynote speech at the International Semantic Web Conference 2005 as an example for the shift from centralized editorial control over Web content by publishers and information syndicators to decentralized control by the user [Wei05].

14.2 Future Directions

The work presented in this thesis opens up several directions for future research.

WIQA Browser. First, it would be interesting to apply the WIQA browser within further application domains. As different domains require different information filtering policies and the representation of different types of quality-related meta-information, this would allow a further evaluation of the WIQA-PL policy language and the Named Graphs data model, and would lead to the development of further extension functions for the WIQA framework. There is already ongoing work in this direction. Radoslaw Oldakowski (Freie Universität Berlin, Germany) is working on applying an extended version of the WIQA browser in an electronic commerce scenario.

WIQA Framework. A second research direction is the integration of the WIQA framework into further web-based information systems. Interesting candidates for being upgraded with information quality filtering capabilities are news portals and newsfeed aggregators, search engines within knowledge management systems, geographical information

systems which render maps with information from multiple sources, and online communities that are used by large numbers of information providers to share information. In general, the WIQA framework can be employed within all types of information systems that capture information together with quality-related meta-information and want to enable information consumers to use this meta-information for information filtering.

A further, more visionary application domain for the WIQA framework is the Semantic Web as a whole. If the vision of the Semantic Web is realized and a growing number of data sources expose their content as RDF, technologies for selecting high quality information from this web of data are needed. Since 2001, the Semantic Web architecture stack [KM01] contains an empty placeholder for a trust layer, which is supposed to support users and software agents in deciding whether to accept or reject information. Along with the Inference Web project, the WIQA framework is the first attempt to fill this placeholder and to propose a solution for one of the unsolved problems on the road to realizing the Semantic Web.

Summing up, this thesis presents an innovative solution to selecting high quality information from the Web. I hope that my work facilitates further research towards the integration of user-centric information quality assessment capabilities into web-based information systems.