The Information Workbench as a Self-Service Platform for Developing Linked Data Applications

Peter Haase, Christian Hütter, Michael Schmidt, and Andreas Schwarte
fluid Operations AG, Walldorf, Germany
[peter.haase, christian.huetter, michael.schmidt, andreas.schwarte]@fluidops.com

ABSTRACT

The Information Workbench is a self-service platform for developing Linked Data applications in the enterprise. Targeting the full life-cycle of Linked Data applications, it facilitates the integration and processing of Linked Data following a Data-as-a-Service paradigm. UI development is based on Semantic Wiki technologies, combined with a large set of predefined widgets for data access, navigation and exploration, visualization, analytics, as well as data mashups with external data sources. In this paper, we present how the Information Workbench can be used to rapidly build industrial-strength Linked Data applications.

Keywords

Self-Service Platform, Linked Data, Visualization

1. INTRODUCTION

In recent years, a large amount of Linked Open Data (LOD) has been published on the Web, comprising over 200 data sets from domains such as media, geography, publication, government, life-sciences, as well as cross-domain data [1]. Growing in size and domain coverage, this data becomes more and more interesting for building innovative applications that integrate heterogeneous data from different sources, in order to overcome the limitations of traditional data management systems. Apart from a new era of Web applications that exploit the large corpus of LOD, this development also offers opportunities in building novel applications for the enterprise by bringing together company-internal data sources with external data, in order to augment and contextualize internal knowledge bases [2].

The development of specific applications that benefit from Linked Data often remains a time-consuming and costly task. First, at the data integration and management side, developers are faced with a variety of new data formats and query languages (such as RDF, OWL, and SPARQL). They also struggle with heterogeneity at data level (facing Linked Data available via HTTP lookups, RDF dumps, and SPARQL endpoints), which requires various new database systems and tools to store, process, and access this data. Second, once the relevant data has been identified and integrated into the system, Linked Data applications require new data interaction paradigms to deal with the specific challenges – and opportunities – of the underlying data formats, such as schema flexibility and data semantics. In particular, to leverage the benefits of Linked Open Data requires the dynamic discovery of available data sources, seamless integration of Linked Data from multiple sources, provenance and information quality assessment. Third, end-user interfaces that implement generic visualization, exploration, and interaction paradigms for Linked Data are important aspects when building Linked Data applications.

Pursuing the goal to lower the entry barrier into the world of Linked Data and to leverage its benefits, the Information Workbench[1] is an open platform for Linked Data applications in the enterprise. It accelerates the development process by abstracting from the technical details behind application deployment, data management, and UI customization. In this paper we present how the Information Workbench can be used as a self-service platform to develop Linked Data applications.

2. THE INFORMATION WORKBENCH

The Information Workbench supports the Linked Data application development process, ranging from data discovery and integration to UI building. Figure 1 shows a high-level architecture diagram. Starting at the bottom, data integration is supported by three complementary concepts:

- For each (internal or external) data source, a so-called *data provider* gathers information from the source, converts it into RDF, and materializes the RDF output in a central data store. Besides built-in mechanisms to integrate semantic data formats such as RDF dumps or data from SPARQL endpoints, the Information Workbench contains generic providers supporting the fast integration of legacy data sources such as relational databases, spreadsheets, Web data accessible through SOAP or REST, and enterprise-internal systems such as LDAP or ERP systems.
- Data can be manually imported into the platform using predefined UIs and APIs such as data upload forms or an interactive Command Line Interface (CLI). Supporting the integration of tabular data and spread-

[1] `http://www.fluidops.com/information-workbench/`

WWW 2012 Developer Track, April 18-20, 2012, Lyon, France

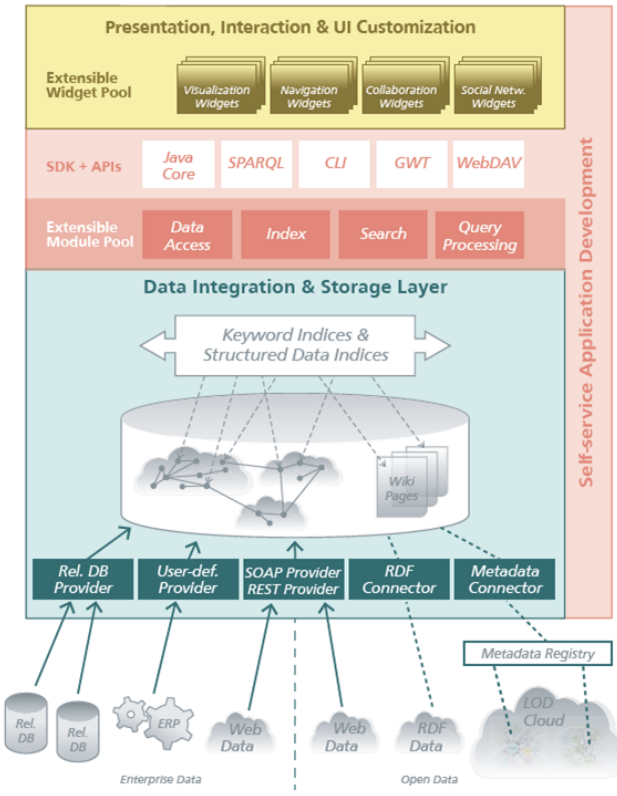

Figure 1: Information Workbench Architecture

sheets, the Information Workbench also offers interfaces to Google Refine, where users can interactively develop sophisticated data mappings.

- Finally, the Information Workbench supports virtualized data integration, where local or public Linked Data sources (such as SPARQL endpoints) can be connected through a federation layer [4], i.e. without materializing the data in the central store.

Once the data has been integrated, every resource in the data graph is automatically associated with a Semantic Wiki page. Semantic Wikis [3] offer built-in mechanisms to access the underlying data, namely to extract and display information from the underlying data graph (e.g., by means of SPARQL queries) and to write directly to the store (e.g., by semantic links that connect resources at data level). The Semantic Wiki thus brings the ability to manage and interlink large amounts of structured and unstructured content imported from existing sources or generated by end users, who can collaboratively annotate, complete, and update content. Accounting for the coexistence of structured and unstructured data, the platform implements advanced search and information access paradigms, ranging from keyword search to complex graph pattern-based search, supporting the user in constructing expressive search queries.

All the core system functionality is extensible and exposed to the outside by different APIs, including Java interfaces, an integrated SPARQL endpoint, or an interactive CLI. Exposing these APIs, the Information Workbench comes with an extensible AJAX-based Web frontend, which supports mashups on both data and UI level, making it possible to interlink data from multiple sources using different visualization and exploration widgets. This *Living UI* enables a homogeneous, personal experience, despite heterogeneous and dynamic underlying data: knowing what, when, and where to show information is realized through an automated and customizable selection of widgets, which implement various paradigms for interacting with the data to exploit the semantics of the underlying data. These widgets can be used to support navigation and exploration, create dashboards for analytics and reporting, collaborative knowledge acquisition, or mash-ups to visualize relationships between data resources, potentially across multiple datasets using available RDF links.

The Information Workbench allows the user to explore data in various views. Besides a standard wiki-style view of entities and other pages, there is a tabular view which details the underlying RDF triples, a graph view which visualizes the RDF graph for the neighborhood of each entity, and a Pivot View for visual exploration.

3. SELF-SERVICE LINKED DATA APPLICATION DEVELOPMENT

In this section, we illustrate the development of an example application using our self-service platform: The *Conference Explorer*[2] is a tool for visitors of a conference. Driven by published metadata related to the conference as well as user-generated content, the Conference Explorer aims to serve as a one-stop-shop for the conference attendee, supporting many activities when planning and attending the conference. Our application presents data grouped around conference events (such as sessions or talks) and people who are associated with those events (such as authors). We also offer a view on the conference as a whole, which does not only serve as an overview and entry point to browse the more specific information, but also provides additional statistical information calculated from the available data. Figure 2 shows the Conference Explorer for WWW 2012.

3.1 Provisioning the Platform as a Service

As an alternative to download the Information Workbench from the fluidOps website, a fully equipped Information Workbench is available as a virtual appliance for immediate deployment from our self-service portal. Following the self-service paradigm, our portal allows wizard-based provisioning of both the application and user-selected data sets in public and private clouds (e.g. Amazon EC2 or VMware vCloud). In our self-service portal users can choose an available Information Workbench template and deploy an instance of the system. We offer templates that are prepopulated with data from various domains[3] such as life science, governmental data, or media data. The data that comes with these instances is connected either via public SPARQL endpoints or through a federation realized on top of a local SPARQL endpoints. Upon completion of the wizard, the Information Workbench instance is deployed in the hosting landscape of the service provider. Once deployed, a system-generated email with access details and credentials will be sent to the user.

[2] `http://conference-explorer.fluidops.net/`

[3] For a list of application areas please see the Information Workbench website.

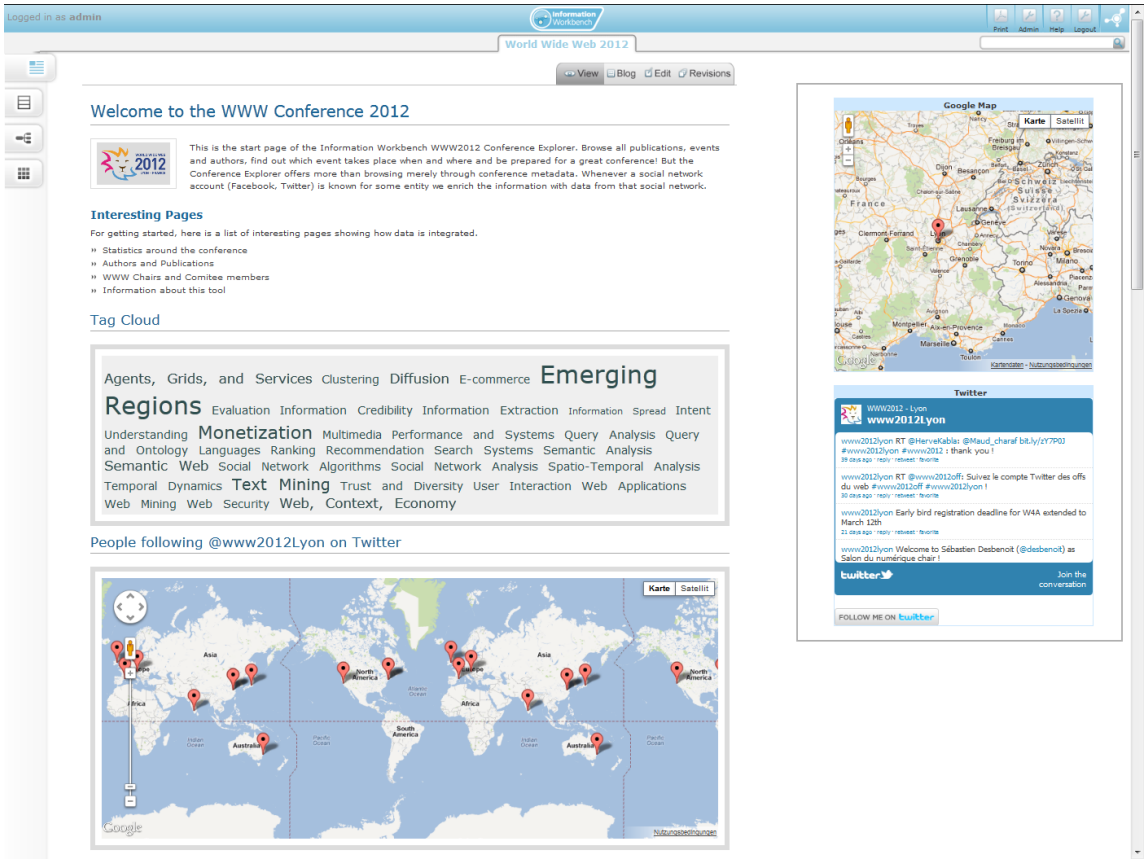

Figure 2: Conference Explorer for WWW 2012

For our example scenario, we select the template for the conference domain, which is configured to use the data from the Semantic Web Conference Corpus as well as the DBLP dataset, which also covers a wide range of conference data. In the selected configuration, the two datasets are included as local RDF databases, i.e. local repositories created from the RDF dumps of the original data sources. The template-based provisioning of the Information Workbench (including the local data repositories) is completed in about 5–10 minutes; the base application is immediately accessible.

3.2 Data Source Discovery

The Information Workbench implements the Data-as-a-Service (DaaS) paradigm. That is, users are able to discover, integrate, and consume available Linked Data ad hoc and on demand. DaaS relies on (1) the availability of individual data sets that can be deployed independently (yet may be interlinked with each other) and (2) the availability of meta information about the content of and access mechanisms to the sources. An integral component of the Information Workbench is a metadata registry, which provides a unified view on metadata and statistics about these data sets by integrating information from different data registries and catalogs such as `http://ckan.org/` (which includes datasets from the Linked Open Data cloud), `http://data.gov/`, and others. The user can explore the metadata catalog as well as the meta information that is available, including domain and size of the data sources, origin, licenses, and information about interlinked data sources.

In our example, we are interested in additional datasets from the social media such as Facebook and Twitter. The data sets are accessible though the respective APIs, for which the Information Workbench contains data providers. Selecting a data set takes us to the detail page, with a description, statistical data, as well as information about the distribution of the data set.

3.3 Data Integration and Deployment

Once a relevant data source has been identified, its data can immediately be integrated into the system. Depending on the access mechanisms that are supported for the data set, the Information Workbench offers options to either load data into the local repository (if an RDF dump is available) or to connect the data virtually through a federation layer (whenever there exists an open SPARQL endpoint). The integration approach is transparent to the end user, which means that (i) within the deployment process the user does not need to be concerned with aspects of physical distribution, access protocols and interfaces, underlying data models etc., and (ii) the details of the integration are hidden at runtime, so both local data and virtually integrated data sources can be queried and accessed in an integrated way.

3.4 Customization of the User Interface

Out of the box, the Information Workbench provides a rich UI, which enables basic interactions with the data as soon as the data has been integrated into the platform. The basic interaction components include tabular and graph-

Figure 3: Pivot View of @www2012lyon's Twitter followers

based visualization and exploration widgets, Semantic Wiki pages for editing and annotating the data, as well as components for semantic and faceted search.

The user interface can then be customized by the users using a large pool of predefined widgets that are shipped with the Information Workbench. These widgets target different data interaction paradigms such as semantic search, data visualization (e.g. as tables, graphs, charts, timelines, maps, etc.), navigation and exploration of the data (e.g. in the form of a graph-based data browser), collaborative editing, knowledge acquisition, as well as mashups with external data sources (such as Youtube, NY Times news feeds, Facebook, and Twitter). In addition, the "Pivot Viewer" shown in Figure 3 allows the visual exploration of the conference's social neighbourhood.

All these widgets can be easily embedded into Semantic Wiki pages and are specified in a fully declarative way using a simple wiki-based syntax. With little effort, the standard views can be customized to create domain- and application specific interfaces. The conference page is based on a simple wiki-based template definition that describes how resources of type 'conference' are presented. The tag cloud in Figure 2, for instance, is generated by the following widget declaration, which builds upon a user-defined SPARQL query:

```
{{
#widget: TagCloud |
  query = 'SELECT ?tag WHERE { ?p foaf:topic ?tag .
  ?p semont:isPartOf www2012:proceedings . }' |
  aggregation = 'COUNT' |
  input = 'tag' |
  output = 'tag'
}}
```

Following the self-service paradigm, our platform also provides *wizards* that support the configuration of widgets, allowing non-expert users to build UIs without the need to specify wiki syntax manually. Furthermore, parameterized query patterns enable users to obtain insights into the data without writing SPARQL queries.

3.5 Extending the Platform via the SDK

While all the steps described previously can be realized without any programming skills, the platform can be extended with own components. Using defined APIs, expert users can easily extend the system functionality by implementing application-specific data providers. The Information Workbench is available in Open Source. Going beyond the current set of features, our easy-to-use SDK can be used to develop and integrate new components such as special-purpose widgets or mashups with only little effort.

4. REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data – The Story So Far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[2] M. Hausenblas. Exploiting linked data to build web applications. *IEEE Internet Computing*, 13(4), 2009.

[3] M. Krötzsch, D. Vrandecic, and M. Völkel. Semantic mediawiki. In *The Semantic Web - ISWC 2006*, pages 935–942, 2006.

[4] A. Schwarte, P. Haase, K. Hose, R. Schenkel, and M. Schmidt. Fedx: Optimization techniques for federated query processing on linked data. In *The Semantic Web – ISWC 2011*, pages 601–616, 2011.