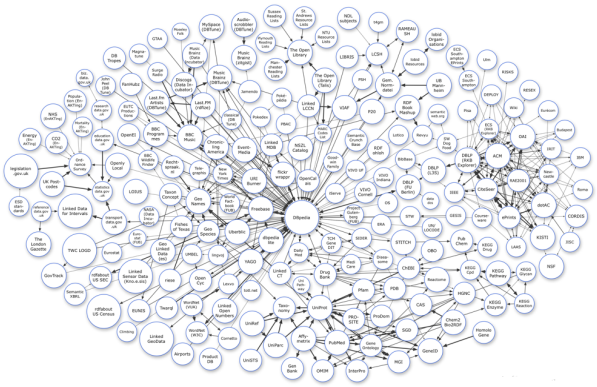
University of Zurich
Department of Informatics

Dynamic and Distributed
Information Systems

Bachelor Thesis September 19, 2011

OGD ZH - A Prototype Implemention



Mengia Zollinger
Zurich, Switzerland

Student-ID: 08-715-427
zollinger.mengia@gmail.com

Advisor: Cosmin Basca

Prof. Abraham Bernstein, PhD
Department of Informatics
University of Zurich
http://www.ifi.uzh.ch/ddis

Acknowledgements

This bachelor thesis would not have been possible without the extraordinary support of many people. Prof. Bernstein, the supervisor of this thesis provided a motivating and critical atmosphere and the adviser Cosmin Basca offered a great support regardless of the time of day. The author would like to thank all the contributors from the project eZürich[1] , especially Georg Andersson from Department of Geomatics and Measurement[2] and Michael Grüebler from Department of Statistics[3] which provided me a great support with the accumulation of data. Additionally, I would like to thank my brother Jan Zollinger who proofreads my bachelor thesis and all my friends and fellow sufferers from the student room.

[1] eZürich: http://www.ezuerich.ch/
[2] Department of Geomatics and Measurement: http://www.stadt-zuerich.ch/geoz
[3] Department of Statistics: http://www.statistik.zh.ch/internet/justiz_inneres/statistik/de/home.html

Abstract

This thesis describes the implementation of a prototype OGD for the city of Zurich. The main focus was the achievement of a data catalogue and several apps for example data visualization. At the beginning, an overview of the procedure and the used framework are introduced, followed by the explanation of the implementation and the resulted challenges. The thesis ends with a comparison of the prototype with similar projects of different countries and with another framework. It is shown that an OGD ZH is possible, but that there are still unsolved issues such as the realization of version control, multilingualism and the automatic generation and assignment of metadata.

Zusammenfassung

Die vorliegende Bachelorarbeit beschreibt die Entwicklung einer Prototyp-OGD-Plattform für die Stadt Zürich. Der Hauptfokus dabei war die Erstellung eines Datenkatalogs und einiger Beispielapplikationen zur Visualisierung der Daten. Zu Beginn wird ein Überblick über die Ziele und die Vorgehensweise der Arbeit gewährt, gefolgt von einer Einführung des verwendeten Frameworks. Anschliessend wird die Umsetzung und die daraus resultierenden Herausforderungen erläutert. Zum Schluss wird der Prototyp mit vergleichbaren Projekten von verschiedenen Ländern und mit einem gleichartigen Framework verglichen. Es wird gezeigt, dass die Umsetzung einer OGD ZH Plattform möglich ist, dass jedoch noch einige Sachverhalte offen sind, wie zum Beispiel die Umsetzung einer Versionskontrolle, die in der Schweiz wichtige Mehrsprachigkeit und die automatische Generierung und Zuordnung von Metadaten.

Table of Contents

1

Introduction

Sharing government information is one of the key elements of a democratic state: citizens should have access to information pertaining public services, basis for decision making during ballots/elections, and information about how their taxes are being used [Bernstein, 2011]. Thereby, transparancy, participation and collaboration are the main issues of the integration of citizens in the paradigm of Open Government [Parycek et al., 2011]. This brings about advantages in transparency, supports innovation processes and helps to improve data quality since citizens are able to provide feedback [Shadbolt, 2010]. On this account the government of Zurich revised the cantonal constitution and added the article 17 which explicitly specified the right of access to official documents [Verfassungsrat, 2008]. Currently, there is a lot of information available but distributed over many separated department websites. A comprehensive data portal can provide a single point of access for public sector data.

Apart from that, the quality of published data varies heavily. Extracting information out of a PDF is much more complex than out of an Excel sheet. Following Tim Berners-Lee model [Heath and Bizer, 2011], published datasets can be classified into five levels of publication quality. At the lowest level, data is available with an unspecified format but with open license. For the next higher level, data is available as machine-readable structured data like Microsoft Excel instead of a PDF or an image. At the next level, data is available in a non-proprietary format like CSV. To rise to the next level data has to be compliant to open standards from W3C such as RDF and SPARQL. To reach the highest level of publication quality, datasets include outgoing links to provide context information.[Heath and Bizer, 2011]

From this initial situation resulted the idea of implementing a prototype containing a registry with open government data. The prototype should include a searchable catalogue with data sources of the city of Zurich. The data source formats shall be selected carefully keeping in mind Tim Berners-Lee concept of cataloging data sources. For illustration purposes one or more apps should be implemented to demonstrate an example usage of data sources. Furthermore, the prototype should be implemented in regard to already deployed similar systems in different countries like in UK or Austria. [Bernstein, 2011]

2

Prototype Implementation

The purpose of this chapter is to outline the scope of the prototype and the process of its implementation. Section 2.3 presents a brief discussion of CKAN[1], an open source software from Open Knowledge Foundation[2] to publish data. This is followed by an outline of the implementation of the data catalogue based on CKAN and the realization of two apps and the resulting challenges during the implementation. The chapter closes with a comparison of the prototype with other CKAN implementations from different countries and a comparison with the commercial data portal software Socrata[3].

2.1 Goal and Scope

The goal of this thesis is to implement a prototype OGD for the city of Zurich. The prototype should serve as proof of concept and demonstrate a possible implementation and visualization of the decisions made. The prototype shall provide a searchable catalogue of static data sources, which can be accessed by different browsers to request data in a user-friendly format. Each data source shall be identifiable by a unique and meaningful URI and should have a transparent and standardized term of use. Employees of the city of Zurich should be able to maintain the contents of the directory. A selection of the static data sources from the government of the city of Zurich should be available in the platform and a sub-selection of these files shall be in a non-proprietary format. The user should be able to request the data required. For illustration issues, especially in relation to the presentation on the Follow Up Meeting of eZürich, one or more example apps shall be implemented using tools such as Exhibit, Google Maps, etc. [Bernstein, 2011]

2.2 Procedure

The prototype was implemented according the following process:

The general scope of the project and the prototype was determined in several meetings with representatives of all participating companies. Since the usage of the data sources is unknown, the files should be available as raw data instead of predefined services. To avoid confusion and frustration, the catalogue should present only available data sources and no empty datasets. Furthermore, users should be able to request unavailable data sources. An extension with a ticketing system could be possible in a future development. The data scope covers only non-real-time data. In a further step real-time data like traffic data form public transport should be included.

[1] Comprehensive Knowledge Archive Network (CKAN): http://ckan.org/
[2] Open Knowledge Foundation: http://okfn.org/
[3] Socrata: http://www.socrata.com/

After perusal of over 700 data sources from the Department of Statistics[4] and the Department of Geomatics and Measurement[5] over 100 data sources were chosen to be published as first attempt. The data sources were selected in a manner to cover a wide range of topics with each subject containing several data sources. The usage of this approach is highly beneficial due to the realistic effigy of a productive system. Additionally, the prototype can be used in the forthcoming hackathon[6]. This selection of data sources is grouped into different categories. A rough classification distinguishes between address data and statistical data representing the two data providers. The underlying layer is based on the structure of the data of the Department of Statistics and the productive CKAN implementation of Open Government Data Vienna[7]. The data samples used in the prototype are categorized in sections such as education, population, health, administration, sport and tourism. The individual data sources were named with a human readable, meaningful and unique URI. Furthermore, the address data sources were named using the same pattern as Open Government Data Vienna, such as "topic name" - location, for example "kindergarden - location" for a directory of all addresses of the kindergartens in Zurich.

Following the model of CKAN, Open Government Data Vienna, Open Government Data UK[8] and the requirements of the Department of Statistics a metadata model was created and all address data sources were added with examples of metadata serving as a reference for uploading by the concerned departments. A sample overview over all licenses used in OGD UK is assorted to provide a reference.

The frontend of the prototype was inspired by the design of the website of eZürich[9] including also the most important elements of a normal website like a section called about and partner logos.

In regard to the presentation of the prototype two apps were implemented. Due to the fact that privacy issues didn't allow using data sources about planning permissions, an elaborate app about event permissions was implemented by using the publishing framework for data-rich interactive webpage Simile Exhibit[10]. A second app was implemented to demonstrate the possibility to map JSON data containing only street data with Google Maps coordinates. Previously, the data source had to be enhanced and converted, because of a less meaningful data granularity and a wrong coordinate format. Unfortunately, it was not possible to implement a third app with other than address data due to time limitations. Instead, the app[11] of the Department of Statistic about Migration in Zurich was integrated.

Due to possible Internet access problems at the presentation location, the prototype was migrated from a remote reachable to a local running instance at the end.

[4]Department of Statistics: http://www.statistik.zh.ch/internet/justiz_inneres/statistik/de/home.html
[5]Department of Geomatics and Measurement: http://www.stadtzuerich.ch/geoz
[6]Hackathon: http://makeopendata.ch/
[7]Open Government Data Vienna: http://data.wien.gv.at/
[8]Open Government Data UK: http://data.gov.uk/
[9]eZürich: http://www.ezurich.ch/
[10]Simile Exhibit: http://simile-widgets.org/wiki/Exhibit
[11]Zurich Explorer: http://statistik.stadt-zuerich.ch/Modules/explorer/test/index.html#story=1

2.3 CKAN

CKAN is an Python based open source data portal software and was developed in 2007 by a non-profit organization called Open Knowledge Foundation for the purpose to provide a registry of open datasets. Meanwhile, several clients including the UK, the Dutch and Norwegian government, use it.

The core functionality includes a catalogue system for metadata, which can be accessed either by a web frontend or an API (Figure 2.1). Tagging and grouping functions are supported but a well-engineered concept for version control and hierarchization is still missing. [Pollock, 2007]

As additional function, a storage system can be integrated to enrich the metadata catalogue with local stored downloadable data sources. But neither the manual upload nor the API are on a sophisticated level. Both alternatives are inconvenient and thus the mime type gets lost during the upload, the file type is indistinguishable during the download and the download URI is less meaningful. A comprehensive change history provides full transparency and a well-established access control mechanism supports the required flexibility. Furthermore, integration with third-party CMS like Drupal is possible. [Pollock, 2007]

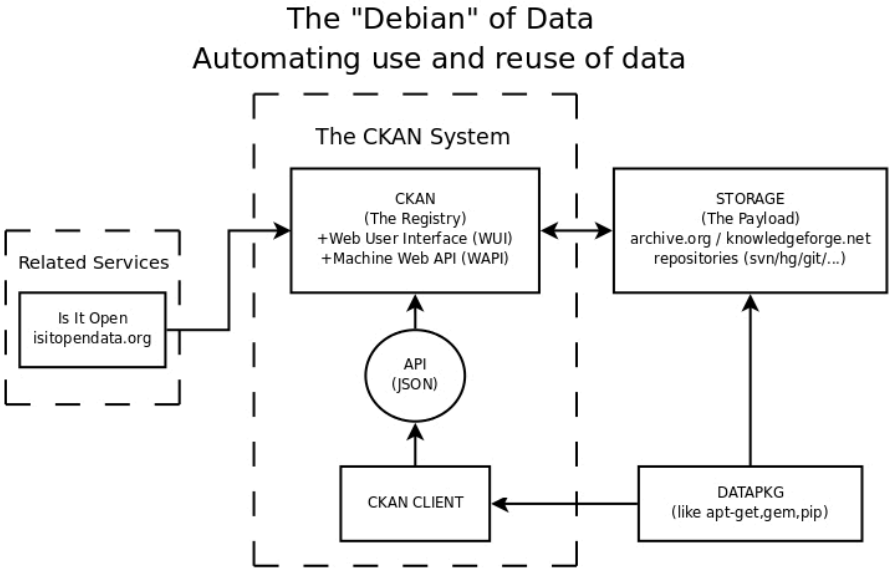


Figure 2.1: Abstract Architecture of CKAN[12]

[12]Source: http://wiki.ckan.net/Purpose

2.4 Implementation

In the following section the implementation of the prototype will be explained in more detail. At the beginning the focus relies on the structure of the general prototype, afterwards the catalogue and the app parts are explained separately. The frontend is divided into four sections. A home screen welcomes new user who would like to browse through the data sets, followed by the data and app parts. The about part covers parties involved, the project vision and possible data topics. The design closly follows the website of eZürich[13], the current driver of this project, mentioning also the University of Zurich and the city of Zurich with their respective logos. By logging in at the backend, the users can edit user profiles, create new revisions and add user groups depending on their respective authorization.

2.4.1 Catalogue

The catalogue provides an overview of all chosen datasets with a short description, license and format information of each download (Figure 2.2). Each dataset is assigned to a unique and meaningful URI and to one or more categories, so the user can filter them and display datasets to a specific topic. Additionally to the datapkg API[14], all common browsers can access the catalogue, returning data in human readable format. Thereby, the employees of the city of Zurich can add, edit and delete data sources and metadata manually or via the API . Either a full-text search or an structured search is possible, for example "Category:sport" returns six results, all assigned to the category "sport".

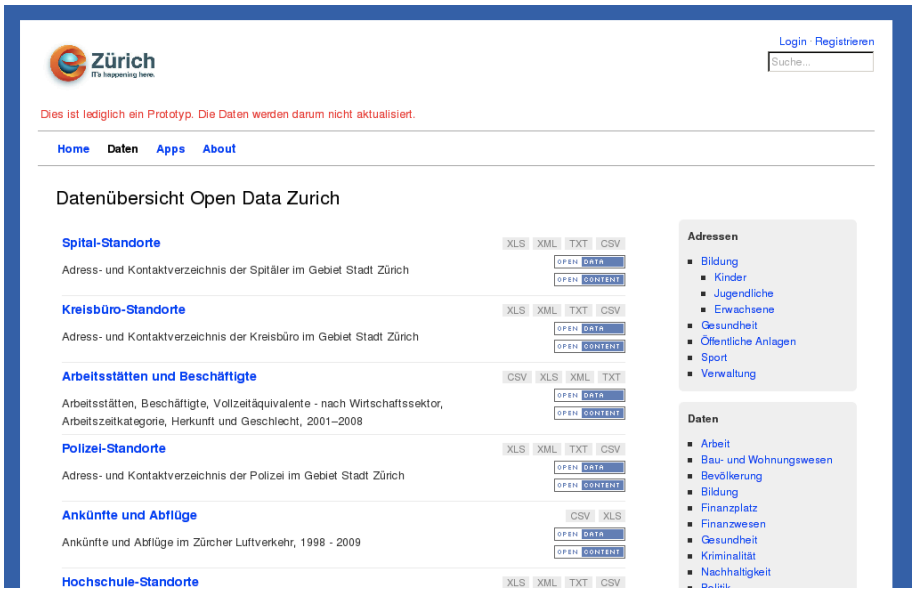


Figure 2.2: Data Screen of the Prototype[15]

Browsing through the content of a specific dataset will lead to the download and the metadata of the dataset. Following the five star methodology of Tim Berners-Lee, data is available with

[13]eZürich: http://www.ezurich.ch/
[14]Datapkg API: http://pypi.python.org/pypi/datapkg/
[15]Source: own illustration

open license, in machine-readable structured data like Microsoft Excel and in non-proprietary formats like CSV, TXT and XML. Thus, the most data provided in the catalogue reaches the third level of the methodology. Due to lack of time, no higher level could be achieved. The metadata contains the following elements using the example of workplace and employees:

Title Arbeitsstätten und Beschäftigte

Name arbeitsstatten-und-beschaftigte

URL -

Notes Arbeitsstätten, Beschäftigte, Vollzeitäquivalente - nach Wirtschaftssektor, Arbeitszeitkategorie, Herkunft und Geschlecht, 2001-2008

Licence OKD Compliant:: Creative Commons Attribution

Tags arbeitsstaetten beschaeftigte betriebe statistik

Category arbeit

URL http://od.ifi.uzh.ch/storage/f/file/bb9b2d16-b7b6-499c-97e0-32863ed54c0d

Format XLS

Description Microsoft Excel

URL http://od.ifi.uzh.ch/storage/f/file/3cb6fb66-c908-4ec9-9490-841f73b2b76f

Format CSV

Description Comma-Separated Values

URL http://od.ifi.uzh.ch/storage/f/file/1580e5cf-e327-41b8-927b-40f3e0c56581

Format XML

Description Microsoft Excel 2004 XML

URL http://od.ifi.uzh.ch/storage/f/file/669db7aa-9e94-4ac1-bd9e-3553e8259b0f

Format TXT

Description UTF-16 Unicode-Text

Group statistik-stadt-zuerich

Author Statistik Stadt Zürich

Author email statistik@zuerich.ch

Maintainer Statistik Stadt Zürich

Maintainer email statistik@zuerich.ch

Update Date 01.09.2011

Update Frequency yearly

Data Type -

Source BFS, Sektion Unternehmensstruktur, Betriebszählungen

Version 1.0 (2009)

State active

The metadata "name" serves as a unique and meaningful identifier for the dataset, defining also its URI. The first URI leads to the description of the data source, whereas the second one leads to the data source itself. In this example, no description is available. Data provider can choose a license from a wide variety of license agreements, but should prefer open licenses such as "Creative Commons" if possible. The individual dataset will lead to similar datasets connected by carefully selected tags. Each dataset can be assigned to a group, which can edit all assigned datasets. Users can also request new datasets by pressing a request button. Each dataset should provide several tags to establish a connection to the other datasets and to support an intuitive browsing behavior for users. To build a catalogue with different topics and to generate a hierarchy between the different datasets, datasets can be assigned to one or more categories. Unfortunately, CKAN does not provide meaningful URI's to the data sources. But following Tim Berners-Lee, URIs should be used as names for things and provide useful information, thus the published data becomes part of a single global data space [Heath and Bizer, 2011]. Consequently, this concept must necessarily be conceptualized in a productive system in order to provide a user-friendly browsing behavior. Additionally, the mime type of the data source gets lost during the upload. In a future release this should be avoided since it is confusing for the data provider when the uploaded file is named with another name than the original one and does not provide the possibility to change the potentially inconvenient, automatically generated name. In addition, while downloading the data source the user would expect that the mime type is recognized and the file opens in the respective appropriate application. If these two functionalities are not supported, it does not correspond with the usability design pattern of learnability. Thus the operation is neither predictable because the user would expect a different reaction of the system, nor does it conform with the principle of familiarity according to which the user would expect that he can transfer his existing uploading and downloading knowledge from different domains to this new system. [Dix, 2004]

Furthermore, the separation of adding metadata and data sources is also inconvenient. From the software provider's point of view, this understandably make sense, since the basic concept of CKAN is to provide a registry for metadata and the local storage of data sources is just an additional feature and introduced in an extension. However, from a data provider perspective adding metadata and data sources separately is conceptually inappropriate. For a future, more user-friendly approach these features should be provided as one entity in the core system.

A large amount of different data source formats facilitates the data access and a comprehensive collection of metadata increases transparency and simplifies the interpretation of data. Additional to the here presented metadata, an accurate description of the values in the data sources should also be provided in the future to avoid interpretation mistakes. An automatic generation of metadata would also be desirable.

In addition to the presented, fix-defined metadata keys, data provider can assign further key-value pairs of metadata. This approach correspond with the human computer interaction principle of adaptability of systems [Dix, 2004].

2.4.2 Apps

There are two apps implemented during this thesis. The first one is an elaborate app about event permissions and the second is a directory of all schools in Zurich. Both apps were implemented by using the publishing framework for data-rich interactive webpages Simile Exhibit [Huynh, 2003] and the integration of Google Maps.

Event Permission App

The first app presents the current event permissions of the city of Zurich. All events can be viewed either on a map, a timeline view or a table view. The data can be filtered by various criteria such as a search box, time span, location, event type or a cloud service. Filtering in one of the views will automatically adapt the representation of data in the other views, making browsing more pleasant. In the map (Figure 2.3) and the timeline view (Figure 2.4) detailed information about an event is being displayed. The underlying data source can be downloaded in both apps.

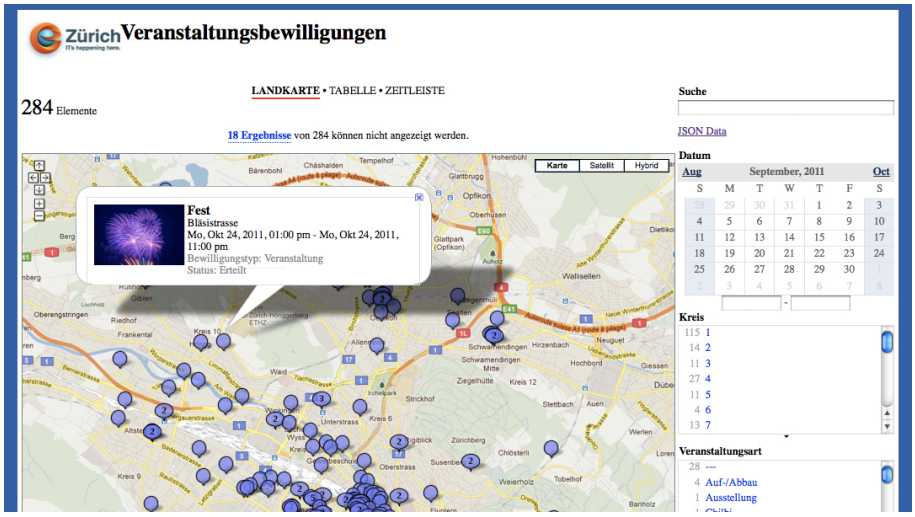


Figure 2.3: Map View of the Event Permission App[16]

School Directory

The second app about the schools based on a similar approach as the event permission app. On a map view the schools of Zurich are displayed. They can be filtered either by school type, location or downloaded as raw source. The particular task of this app was to demonstrate the possibility of mapping JSON data containing only street data with Google Maps coordinates.

[16]Source: own illustration
[17]Source: own illustration

Figure 2.4: Timeline View of the Event Permission App[17]

2.4.3 Challenges

Various factors posed significant constraints to the project. The most important challenge was the short span of time of the project in which a considerable amount of data had to be reviewed and categorized. Also, the factor that the project was conducted during the holiday season impeded collaboration with the project contributors since they were partly absent. The limited amount of information on the other hand on the web about CKAN was predominately compensable by the elaborate documentation and the mailing list of the CKAN team.

Since CKAN is a relatively new framework its core functionality is still not entirely developed. Despite the fact that additional functions can be added by extensions, its functionality is not very convenient, which is shown in the upload and storage process of data sources. Additional core functionality such as the possibility of versioning data has to be introduced by self-implementation. Furthermore, the underlying unknown structure, which is based on Python and integrates several frameworks such as Pylons[18], Pairtree[19], Paste[20] and Datapkg[21], posed another significant challenge for the author.

The main challenge concerning the apps section was that the underlying data sources were not entirely meaningful, which is why they first of all had to be enhanced. For example, the coordinates were expressed in the Swiss format and had to be converted to the format of Google Maps and the street data had to be assigned to the corresponding quarter to allow filtering by category. Furthermore, the framework Simile Exhibit[22] was at the beginning unknown for the author.

[18]Python Pylons: https://www.pylonsproject.org/
[19]Python Pairtree: http://pypi.python.org/pypi/Pairtree#downloads
[20]Python Paste: http://pythonpaste.org/
[21]Datapkg: http://pypi.python.org/pypi/datapkg/
[22]Simile Exhibit: http://simile-widgets.org/wiki/Exhibit

3

Comparison

In the previous chapter CKAN and the implementation of the OGD ZH prototype has been presented followed by the resulted challenges. In the subsequent section the prototype will be compared at first with a selection of other CKAN implementations from several countries and afterwards with the commercial data publishing system Socrata.

3.1 Comparison with other CKAN implementations

CKAN is widely used by several governments including the UK's, Norwegian's and the Dutch government and many other data publishers [Pollock, 2007]. The next section compares the implemented OGD ZH prototype with a selection of already deployed CKAN implementations.

3.1.1 UK

One of the most mature CKAN implementations is with certainty the data portal of the government of UK[1]. It was launched under the direction of the founder of the World Wide Web, Tim Berners-Lee [EU Commission, 2011] and contains over 7400 datasets with a wide range of metadata and a huge amount of apps. Its datasets can be filtered based on various different criteria, for instance based on relevancy, the update timestamp, ratings, resource formats, tags, nations and publishers. Furthermore, it provides a large offer of interactive options with the citizens of the respective country. They can comment datasets, request new data, share ideas or self-implemented apps, participate in forums or inform themselves in blogs. Integration of social networks like Facebook, LinkedIn, Twitter or Google Buzz is possible.

What is truly remarkable at this CKAN implementation is that a version control is implemented by adding data sources with different timestamps to the same dataset. Additionally, different data sources belonging to the same topic are also included in the same dataset. This approach contradicts with the main idea of CKAN, giving each dataset a unique identifier. However, it bypasses the current poorly supported possibility of grouping datasets with similar topics and connects data sources with different versions. Therefore, the data portal of the government of UK provides a considerably larger spectrum of datasets and functionality as the prototype, but recedes from the idea of uniquely identifiable datasets. [UK, 2010]

[1] OGD UK: http://data.gov.uk/

3.1.2 Austria

The government of Vienna conducts also an open government data platform[2], providing dataset, apps and information about open data events. The major area of interests lies here on the reference to the website of the government, which provides a conspicuous user centered approach offering direct access to online forms for example to order a new passport or change the own name. Apart from that, it contains a similar approach as OGD UK when it comes to providing datasets and apps. Furthermore, it contains a much more precise description of the data sources, mentioning also its attributes with an auxiliary description. [Magistrat Stadt Wien, 2011]

Similar as OGD Vienna which follows the model of the government website of Vienna, the frontend of the prototype OGD ZH was designed after the website of eZürich, which was the main project driver so far. The disadvantage of this approach is, that this website is probably not well-known to the citizens. It would be therefore recommendable to refer to the website of the government website of Zurich while implementing a productive system. Beforehand, it must be determined which government part drives the project, the city or canton of Zurich or if it is a cross-cantonal project.

3.1.3 International Aid Transparency Initiative

In contrast to the above mentioned platforms, the International Aid Transparency Initiative[3] does not contain a wide spectrum of datasets but it is specialized on one specific topic, publishing data about international development activities. Additionally, IATI provides a single point of access to metadata and links to datasets hosted by donor agencies while not hosting data itself, whereas similar CKAN implementations, including the OGD ZH prototype, store most of their datasets locally.

In addition, IATI concentrates on one data format, XML, and one metadata format, JSON, providing a homogenous access to the data sources, whereas comparable systems provide a wide variety of formats to support its user as good as possible. OGD ZH provides currently the following data formats: XLS, XML, CSV and TXT. [IATI, 2005]

3.2 Comparison with another data portal system

The implemented prototype is based on CKAN, an open source solution of Open Knowledge Foundation. In the following section the commercial solution Socrata is presented and compared with CKAN.

3.2.1 Socrata

Socrata[4] provides a commercial, cloud-based open data solution for static and real-time data, presented in different views as embedded tables, visualizations or various downloads. The main customers are American governments such as the USA's[5], the State of Oregon[6], the State of Okla-

[2]Open Government Data Vienna: http://data.wien.gv.at/
[3]IATI: http://iatiregistry.org/
[4]Socrata: http://www.socrata.com/
[5]OGD US: http://www.data.gov/
[6]OGD Oregon: http://data.oregon.gov/

Type	Description	OGD ZH	OGD UK	ODG Vienna	IATI
Content	Topic	Government data of Zurich	Government data of UK	Government data of Vienna	International development activities
	Amount of datasets	50	7400	60	270
	Data formats	XLS, CSV, XML, TXT	XLS, CSV, XML, ASP, RDF, HTML, PDF, TXT	CSV, XML	XML
	Local hosting	Yes	No	Yes	No
	Hierarchization	Constrained, multiple layers	Dataset contains similar topics of data sources	Constrained, one layer	No
	Version control	Constrained	Dataset contains several version of data sources	No	Constrained
	Quality control	Email	Comment, Rating	Forum	Email
	Metadata	General description of dataset	General description of dataset	Description of particular data source	General description of dataset
	Design	Similar as eZürich	Similar as HM Government	Similar as Government of Vienna	Similar as the IATI website
	Web accessability/Multilingualism	No/No	No/No	Visual support/No	No/No
Interaction	Filter	Category, name, title, notes, tags, publisher	Category, name, title, notes, tags, relevance, update timestamp, rating, resource format, nation, publisher, recommendation	Category, name, title, tags	Source, publisher, country, title
	User interaction	Download	Download, comment, share ideas and apps, forum, blog	Online forms, forum, sharing of apps	Download
	Social network integration	No	Facebook, Twitter, LinkedIn, Google Buzz	Twitter	No
	Visualization	Apps	Apps	Apps	No
	Request dataset	Yes	Yes	Yes	No

Table 3.1: Comparison of the Prototype with other CKAN implementations

homa[7] and the cities of Chicago[8] and Seattle [9]. [Socrata, 2007] Socrata offers an extensive range of core functions, which is why the United States Government could realize a first version of their platform within 120 days [Houghton and Garnar, 2011].

Socrata supports the most important data formats such as JSON, XML, RDF, XLS, XLSX, CSV, PDF and TXT. A great benefit of Socrata in comparison with CKAN is the integration of features to visualize data directly on the platform. Various visualization tools like filters, maps, calendar and charts are at free disposal and offer a intuitive usage. The resulting views are presented in direct relation to the underlying data source, allowing full transparency. The generated views can be downloaded easily, sent as email or embedded in one's own website, Facebook or Twitter account. This can be done either by web frontend or by API. Through this easy approach of visualization assistance, the probability that citizens get interested and involved is significantly higher than in CKAN because the users of CKAN tediously have to manually download the raw data, build charts locally and upload them again. This is more time-consuming and takes for granted that the users have already their own visualization tools. Finally, it costs additional effort to upload the visualizations and the visualizations are therefore not connected with the original data which is why changes in the raw data will not automatically be reflected in the respective visualization.

Using Socrata citizens can discuss about datasets, request new ones and monitor their status, rate and discuss the requests. Datasets can also be assigned to categories and topics, but Socrata just supports one level of hierarchy, which is still insufficient because using a large amount of data would not allow a scaling and get quite complex.

It is also possible to present real-time data. The Government of Seattle for example informs the citizens about 911 fire calls [Seattle, 2011] and updates the underlying data source every ten minutes.

Two important but apparently not yet implemented features are the version control and the support of multilingualism. CKAN doesn't support these functions yet either. Therefore, Socrata is deployed mainly in the English-speaking world, it is not probable that multilingualism will be supported soon, whereas CKAN is used in the multilingual Europe. [Socrata, 2007]

Taking all these aspects into consideration, it can be stated that the commercial Socrata supports a wider spectrum of functions but has also lack of main functions such as version control and multilingualism and the data storage is problematic for privacy reasons. But the approach of providing tools to directly visualize data on the platform, tracking the popularity of data and the possibility to embed generated views in social networks offers an extraordinary additional value and can lead to an increased involvement of the citizens. Therefore, it would be advisable to analyze in a first phase of a software evaluation[10] Socrata and similar systems, giving particular focus to their handling of the human computer interaction and their functions offered.

[7] OGD Oklahoma: http://data.ok.gov/
[8] OGD Chicago: http://data.cityofchicago.org/
[9] OGD Seattle: http://data.seattle.gov/
[10] From the requirement analysis to the creation of a shortlist; after the SPEED method[Mogicato, 2011]

Type	Description	CKAN	Socrata
System	Language	Python	Unknown
	Source code availability	Open source	Commercial, software as a service
	Distribution	Europe	America
	Embedding	Extension: drupal, phpBB, etc	Unknown
Content	Hosting	Extension: local, mime type loss	Cloud based, depending on service level
	Hierarchization	Multiple layers	One layer
	Version control	No	Yes
	Data formats	Undefined	Depending on service level
	Update rate	Static	Static, real time
	Quality control	No	Certification, user rating, data rating, report datasets
	Performance metrics	Extension: Google Analytics	Own and Google Analytics
	Group, tag, filter	Yes	Yes
	Custom metadata	Yes	Depending on service level
Interaction	Access	Frontend, API	Frontend, limited API calls
	Search	Selective metadata	All metadata and table entries
	Social network	No	Facebook, Twitter, embeds in websites
	Multilingualism	Partially	No
	Dataset comparison	No	No
	Real time interaction	No	Comments, rating
	Non-real time interaction	Extension: forum	Yes
	Visualization	No	Maps, calendar, charts, filter, column hiding
	Dataset request	No	Yes

Table 3.2: Comparison of CKAN and Socrata

4

Future Work

An important issue to realize a productive system is the augmentation and improved selection of available data sources. The datasets must be attractive to potential user and cover topics such as health, education, transport, crime, tax revenues and government spending data. Datasets should be assignable to their respective location to enable comparisons, for example of the expenses of cantons Zurich and Berne. This requires a standardization and improvement of metadata. An automatic development of categories and tags and an inclusion of a refinement application like Google Refine[1] would be desirable. In a future step, the metadata should also include a description of the data source attributes. Furthermore, there is a potential target group, which is not interested in raw data to analyze and process, but merely would like to be informed about a specific topic. Thus, an integration of tools to view raw data for example as bar or pie charts would also be helpful to interpret the data based on the model of Socrata.

An enhanced integration of the ideas of Linked Data would also be desirable. Thus, the dataset section should be extended by more formats, for example RDF and OData[2] and include open vocabularies such as FOAF[3], SKOS[4], SIOC[5], Dublin Core[6] and DOAP[7].

A productive system needs also a concept of maintenance and version control. The enlargement beyond static data to live-streaming data, for example data of public transport would also be an interesting topic. This could be particularly interesting since there is no other country which has realized this subject yet.

An integration of an external open source software like the content management system Drupal[8] or the forum software phpBB[9] would be useful. Multi-language support in all four native languages extended with English would also be helpful but challenging. The quality of datasets could be measured in two different approaches. On the one side, they shall be assessed following the five star methodology [Heath and Bizer, 2011] outlined by Tim Berners-Lee. On the other side, users should be able to rate them to estimate the relevance and usefulness and report inappropriate contributions. Consequently, the most popular datasets could be displayed in the main page. Popular concepts from forums could be adopted. Users can graduate by writing comments or adding well-rated datasets. A recommendation service along the lines of OGD UK[10] which recommends the users similar datasets might be possible. All this measures could help building a community around the Open Government Data.

[1] Google Refine: http://code.google.com/p/google-refine/
[2] Open Data Protocol: http://www.odata.org/
[3] Friend Of A Friend Project: www.foaf-project.org
[4] Simple Knowledge Organization System: www.w3.org/2004/02/skos
[5] SIOC Project: http://sioc-project.org/
[6] Dublin Core Metadata Initiative: http://dublincore.org/documents/dces/
[7] Description of a Project: http://trac.usefulinc.com/doap
[8] Drupal - An Open Source CMS: http://drupal.org/
[9] PHP Bulletin Board: http://www.phpbb.com/
[10] Open Government Data UK: http://data.gov.uk/

Following the approach of OGD Vienna[11] an integration of a virtual administration could also be realized, thus users can directly access documents and forms to request services like new passports or residence requests.

Furthermore, an interconnection with other OGD should be realized by adding references to the data portal, adding the data portal in open data catalogue lists[12] like this of the Open Knowledge Foundation or integrate it in the OGD EU portal[13] which provides a single point of access to national, regional and local OGDs throughout Europe.

But beforehand, many legal, privacy and political aspects such as license agreements, data and project scope and financial issues such as cost of maintenance and loss of revenue from departments have to be determined before this project can be implemented on a large-scale.

[11]Open Government Data Vienna: http://data.wien.gv.at/
[12]List of European Open Data Catalogues: http://lod2.okfn.org/eu-data-catalogues/
[13]Open Government Data EU: http://publicdata.eu/

5

Conclusions

The goal of this bachelor thesis was to implement a prototype OGD for the city of Zurich, which includes a searchable catalogue of data sources and several apps to illustrate potential visualizations. This goal was accomplished by the implementation of the prototype Open Data Zurich described in section 2.4 of this thesis. There are some unsolved issues such as the missing possibility of version control, the missing support of multilingualism and some constraints in connecting similar datasets to a specific topic. Future enhancements could also include community functions such as wikis and forums to demonstrate and discuss best practices and the possibility to comment and rate datasets. Also, a login option with a user availability status, chat and Skype options could be options worth implementing in a productive system. While there is no clearly need of a complete new social network it would be beneficial to connect the application with existing social networks and to implement the ability to enrich them with datasets, visualization and apps.

Due to the fact that there are various implementation efforts of Open Government data taking place all over the world the government of Zurich should closely monitor and analyze those different approaches and their respective success. Based on this and on their local requirements they can then customize and implement their own software. However, an isolated solution for the city of Zurich should be avoided since interested companies may want to contribute information and integrate more services and other cantons may be would like a cross-cantonal solution.

Because of the paragraph [Verfassungsrat, 2008] about transparency in the cantonal constitution the legal need for an OGD for the city of Zurich is clearly given. In addition, in our competitive contemporary business environment it is increasingly important to encourage innovation by providing open information to the citizens of a government. This may help to improve government processes, increase transparency, impede corruption and is vital in the positioning of Zurich as one of the top location for ICT companies. Since the realization of such a project is a balancing act between transparency and privacy a strong policy must be elaborated that includes the maintenance and licensing concept as well as a control mechanisms to keep the data quality and the contribution of external users high. Consecutively, the table 5.1 contains a non-exhaustive list of open tasks for the implementation of a productive system. Last but not least, the open access to information is one of the most important aspects of a working democracy to facilitate the veridical formation of opinion of the citizens.

Type	Name	Description	Must/Should be
Administrative	Project driver	Who is the driver of the project?	M
	Project scope	How is the project scope defined?	M
	Interconnection	How should the project be interconnected with other projects?	M
	Funding	How is the project funded?	M
	Legal	Which legal actions does it need before the project start?	M
	Data	Which data and metadata should be included?	M
	License	Under which license can user reuse data?	M
	Privacy	How should a privacy policy look like?	M
	Formats	Which data formats should be provided?	S
	Design	Which design should the frontend have?	S
Technology	Security	How is the security concept defined?	M
	System	Make or Buy? Which framework?	M
	Data upload	How to upload data automatically?	M
	Hierarchization	How can data be categorized and structured hierarchically?	S
	Version control	How can version control be supplied?	S
	Geo mapping	How can data be geo mapped?	S
	Static / real time data	How can real time data be supported?	S
	Refinement Application	How can refinement tools such as Google Refine be integrated?	M
	Performance metrics	How can analytic systems such as Google Analytics be integrated?	S
	Linked data	How can concepts from linked data be integrated?	S
	Ticketing System	How to deal with requests for new datasets?	M
	Multilingualism	Which languages should be supported?	S
	Visualization	How should visualization of data be supported (maps, chars, etc)?	S
	Social Network	How can common social networks be integrated?	S
	User interaction	How can user interation be supported (forums, blogs, charts, Skype)?	S
T, A	Maintenance	How should a maintenance concept look like?	M
	Government services	How can government services be embedded?	S
	Quality control	How can the quality of the contribution be ensured?	M

Table 5.1: Task list for the implementation of a productive system

A

Appendix

A.1 Attachments

The following section contains an overview about the attachments on the enclosed CD:

Bachelor Thesis A complete copy of this bachelor thesis called ”Bachelorarbeit” as PDF file

Summary A summary of this bachelor thesis in German called ”Zusfsg” as a text file

Abstract An abstract of this bachelor thesis in English called ”Abstract” as a text file

Source Code The source code of the CKAN implementation, the event permission app and the school directory app

Data A survey of the selected catalogue data called ”Data” as PDF file

Metadata An overview of the catalogue metadata called ”Metadata” as Excel file

Decisions A summary of all discussions and decisions called ”Decisions” as PDF file

A.2 Source Code of the Event Permission App

As an exemplary demonstration of source code, the event permission app will be discussed in the following section. Due to clearity reasons the code has been strongly cut in the JSON containing only two items representing all other entries. For the original source code, please have a look at the attachment on the CD.

A.2.1 HTML File

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html>
<head>
  <title>Veranstaltungsbewilligungen</title>


  <link rel="exhibit/data" type="application/json" href="veranstaltungsbewilligungen.js">

  <script src="http://api.simile-widgets.org/exhibit/2.2.0/exhibit-api.js" type="text/javascript">
      </script>
  <script src="http://api.simile-widgets.org/exhibit/2.2.0/extensions/time/time-extension.js" type
      ="text/javascript"></script>
  <script src="http://api.simile-widgets.org/exhibit/2.2.0/extensions/map/map-extension.js?gmapkey
      =ABQIAAAAxgQlw_Lp55k6bwRypl0zyBSXSFUY-tcBT_W0NcDa2XqYVNcFFBR-WalSBjsmAl0BfrCAfWrGqO2c8g"></
      script>
  <script src="http://api.simile-widgets.org/exhibit/2.2.0/extensions/calendar/calendar-extension.
      js"></script>
  <style>
body {
margin-left: auto;
margin-right: auto;
padding-top: 10px;
}
h1 { color: #000;
}
table.permission {
border: 1px solid #ddd;
padding: 0.5em;
}
div.name {
font-weight: bold;
font-size: 120%;
}
.location {
}
.time {
}
.relationship {
color: #888;
}
  </style>
</head>


<body style="background-color: rgb(42, 93, 174);">

<table style="background-color: rgb(255, 255, 255);" align="center" border="0" width="80%">

  <tbody>

    <tr>

      <td halign="center">
      <h1 align="justify"><a title="Open Data Zurich Home" href="../index.html"><img style="
          display: inline;" title="Open Data Zurich Logo" alt="Open Data Zurich Logo" src="
          ezuerich_logo.jpg" align="middle" height="80"></a>Veranstaltungsbewilligungen</h1>

      </td>

    </tr>
```

```
55    </tbody><tbody>

57      <tr>

59        <td>
          <table style="background-color: rgb(255, 255, 255);" width="100%">
61
            <tbody>
63
              <tr valign="top">
65
                <td ex:role="viewPanel">
67  <!-- Kachelansicht -->
                <table ex:role="lens" class="permission">
69
                  <tbody>
71
                    <tr>
73
                      <td><img ex:src-content=".IMGURL" height="90"></td>
75
                      <td>
77                    <div ex:content=".VERANSTALTUNGSART" class="name"></div>

79                    <div ex:content=".ORT" class="location"></div>

81                    <div> <span ex:content=".ZEITVON" class="location"></span> - <span ex:content=".
                          ZEITBIS" class="time"></span></div>

83                    <div style="color: rgb(136, 136, 136);">Bewilligungstyp:
                      <span ex:content=".BEZEICHNUNG" class="relationship"></span></div>
85
                      <div style="color: rgb(136, 136, 136);">Status:
87                    <span ex:content=".STATUS" class="relationship"></span></div>

89                    </td>

91                  </tr>

93                </tbody>
                </table>
95
    <!-- Map -->
97              <div ex:role="view" ex:viewclass="Map" ex:latlng=".LAT" ex:zoom="13" ex:color="#7979fc
                     " ex:mapheight="900" ex:center="47.384956,8.536034" ex:eventlabel=".BEZEICHNUNG">
                     </div>

99  <!-- Tabelle -->
                <div ex:role="exhibit-view" ex:viewclass="Exhibit.TabularView" ex:columns=".
                     VERANSTALTUNGSART, .KREIS, .ORT, .HAUSNR, .ZEITVON, .ZEITBIS" ex:columnlabels="
                     Veranstaltung, Kreis, Ort, Hausnummer , von, bis" ex:columnformats="list, list,
                     list, list, list, list" ex:sortcolumn="1" ex:sortascending="true"></div>
101
    <!-- Zeitleiste -->
103             <div ex:timelineheight="350" ex:eventlabel=".BEZEICHNUNG" ex:colorkey=".
                     VERANSTALTUNGSART" ex:start=".ZEITVON" ex:viewclass="Timeline" ex:role="view"></
                     div>

105             </td>

107             <td width="25%">
    <!-- Filter -->
109             <div ex:role="facet" ex:facetclass="TextSearch" ex:facetlabel="Suche"></div>

111             <div>
                <p><a href="veranstaltungsbewilligungen.js" target="_blank">JSON Data</a></p>
113
                </div>
115
                <div id="event_date" class="facet" ex:role="facet" ex:begindate=".ZEITVON" ex:enddate=
                     ".ZEITBIS" ex:facetclass="DatePicker" ex:collapsible="true" ex:facetlabel="Datum"
                     ></div>
117
```

```
            <div ex:role="facet" ex:expression=".KREIS" ex:facetlabel="Kreis" ex:sortmode="value"
                ex:fixedorder="1;2;3;4;5;6;7;8;9;10;11;12" ex:showmissing="false"></div>

            <div ex:role="facet" ex:expression=".VERANSTALTUNGSART" ex:facetlabel="
                Veranstaltungsart" ex:showmissing="false"></div>

<!-- Tag Cloud -->
            <div ex:role="facet" ex:facetclass="Cloud" ex:expression=".VERANSTALTUNGSART" ex:
                height="100px" ex:facetlabel="Cloud" ex:showmissing="false"></div>

            </td>

          </tr>

        </tbody>
      </table>

      </td>

    </tr>

  </tbody>
</table>

</body>
</html>
```

A.2.2 JSON File

```
{ properties: {
            "MYKEY" : {
                valueType: "item"
            }
        },

    types: {
            "Bewilligung" : {
                pluralLabel: "Bewilligungen"
            }
        },
    "items" :       [
        {
            "KURZBEZ" :             "V10",
            "ZEITBIS" :             "2011-08-29T12:00:00-04:00",
            "IMGURL" :              "fest.jpg",
            "OST" :                 680795.206,
            "type" :                "Item",
            "STATUS" :              "Erteilt",
            "DATE" :                "2011, 8",
            "ZEITVON" :             "2011-08-29T08:00:00-04:00",
            "LATCH" :               "252361.67,680795.206",
            "label" :               "513882",
            "LAT" :                 "47.4171038945589,8.50933215558693",
            "VERANSTALTUNGSART" :   "Fest",
            "HAUSNR" :              "495",
            "KREIS" :               "11",
            "NORD" :                252361.67,
            "ORT" :                 "Wehntalerstrasse",
            "MYKEY" :               513882,
            "BEZEICHNUNG" :         "Veranstaltung"
        },

            {
            "KURZBEZ" :             "V32",
            "ZEITBIS" :             "2011-10-22T12:00:00-04:00",
            "IMGURL" :              "demo.jpg",
            "OST" :                 685628.29,
            "type" :                "Item",
            "STATUS" :              "Erteilt",
            "DATE" :                "2011, 10",
            "ZEITVON" :             "2011-10-22T09:00:00-04:00",
            "LATCH" :               "250983.75,685628.29",
            "label" :               "534198",
```

```
                "LAT" :                "47.4041012436496,8.57311854392665",
46              "VERANSTALTUNGSART" : "Info polit.",
                "HAUSNR" :             "524",
48              "KREIS" :              "12",
                "NORD" :               250983.75,
50              "ORT" :                "Winterthurerstrasse (12.3)",
                "MYKEY" :              534198,
52              "BEZEICHNUNG" :        "Standaktion auf \u00F6ff. Grund, politisch"
            }
54      ],
        "properties" : {
56          "ZEITBIS" : {
                "valueType" : "date"
58          },
            "OST" :     {
60              "valueType" : "number"
            },
62          "NORD" :    {
                "valueType" : "number"
64          },
            "ZEITVON" : {
66              "valueType" : "date"
            },
68          "MYKEY" :   {
                "valueType" : "number"
70          }
        }
72  }
```

List of Figures

List of Tables

Bibliography

[Bernstein, 2011] Bernstein, A. (2011). data.zh.ch - A Prototype Implementation, B.Sc. Thesis Description. page 1.

[Dix, 2004] Dix, A. (2004). *Human-Computer Interaction*. Pearson Education.

[EU Commission, 2011] EU Commission (2011). Europe's Public Data. Website. Available online at `http://publicdata.eu/`; visited on September 9th 2011.

[Heath and Bizer, 2011] Heath, T. and Bizer, C. (2011). *Linked Data - Evolving the Web into a Global Data Space*. Morgan & Claypool.

[Houghton and Garnar, 2011] Houghton, V. and Garnar, P. M. L. (2011). Data.gov: The Risks and Benefits of Transparency. page p. 2.

[Huynh, 2003] Huynh, D. (2003). Simile exhibit. Website. Available online at `http://simile-widgets.org/wiki/Exhibit`; visited on August 21th 2011.

[IATI, 2005] IATI (2005). International Aid Transparency Initiative Registry. Website. Available online at `http://iatiregistry.org/`; visited on September 8th 2011.

[Magistrat Stadt Wien, 2011] Magistrat Stadt Wien (2011). Open Government Data - Offene Daten für Wien. Website. Available online at `http://data.wien.gv.at/`; visited on September 8th 2011.

[Mogicato, 2011] Mogicato, R. (2011). Evaluierung von Standardsoftware - Auswahlentscheid. page 28.

[Parycek et al., 2011] Parycek, P., Kripp, M. J., and Edelmann, N. (2011). Proceedings of the International Conference for E-Democracy and Open Government. In *Proceedings of the International Conference for E-Democracy and Open Government*.

[Pollock, 2007] Pollock, R. (2007). Comprehensive Knowledge Archive Network (CKAN). Website. Available online at `http://ckan.org/`; visited on August 31th 2011.

[Seattle, 2011] Seattle (2011). Seattle Real Time Fire 911 Calls. Website. Available online at `http://data.seattle.gov/Fire/Seattle-Real-Time-Fire-911-Calls/kzjm-xkqj`; visited on September 7th 2011.

[Shadbolt, 2010] Shadbolt, N. (2010). Towards a pan EU data portal - data.gov.eu. page 5.

[Socrata, 2007] Socrata (2007). Socrata - the open data company. Website. Available online at `http://www.socrata.com/`; visited on September 8th 2011.

[UK, 2010] UK (2010). Opening up Government. Website. Available online at http://data.gov.uk/; visited on September 5th 2011.

[Verfassungsrat, 2008] Verfassungsrat (2008). Kantonsverfassung - Menschenbild, Werteordnung und StaatsverstŁndnis. Abgerufen am 23.06.2011 unter http://www2.zhlex.zh.ch/appl/zhlex_r.nsf/0/ABF964058B1A5907C12577E10039C7EB/$file/101_27.2.05_71.pdf.