The Kasabi Information Marketplace

Knud Möller
Kasabi
43 Temple Row
B25LS, Birmingham, UK
km@kasabi.com

Leigh Dodds
Kasabi
43 Temple Row
B25LS, Birmingham, UK
ld@kasabi.com

ABSTRACT

Publishing and consuming structured data on the Web is becoming more and more common across domains as varied as the public sector, the media, cultural institutions, the manufacturing industry or retailers. Kasabi, an online data market based on linked data principles, offers data publishers an easy way to publish, link and monetise data, while giving developers of data-centric applications access to this data in different formats and through a number of different interfaces. This short paper introduces Kasabi and gives an overview of its capabilities and the functionality it offers to developers through its APIs.

Keywords

linked data, open data, data market

1. THE WEB OF DATA

Across many different domains, companies and organisations more and more see the benefit in both publishing their own data in a structured format on the Web, as well as making use of existing structured data. Governments and other public sector organisations embrace an open data policy, informing and providing their public with valuable information. One of the latest examples of this development is the announcement of the European Commission to adopt an open data strategy (e.g., [1] or [6]). Media organisations such as the NY Times or the BBC, cultural institutions such as the British or German National Libraries, large companies such as Volkswagen or BestBuy are all making data available in a structured, interlinked format on the Web, creating what is known as the "Web of Data" [4]. Among other things, the availability of this data drives the development of new applications, increases its publisher's visibility, enables new forms of journalism and facilitates integration, co-operation and communication in ways that closed and isolated databases did not allow.

On the consumer side of things, the big search engine providers Google, Yahoo! and Microsoft have taken note of this development and jointly published schema.org, a set of vocabularies for structured data embedded in webpages. This data allows the search engine to tailor search results towards the content of the page, but can just as well be harnessed by any other software accessing the same page.

Publishing data on the Web can be as simple as uploading a CSV file to a server. However, to best facilitate reuse and integration, data should be published according to a set of best practices, including a particular data format (RDF[1]), linking to other datasets, infrastructure to make each data item resolvable in different data formats, interfaces to allow structured queries (using the SPARQL language [7, 5]), dataset descriptions, or entering one's dataset in a registry to improve findability.

As with any new technology that is not yet standard procedure, adhering to these best practices and setting up the necessary infrastructure can present a significant challenge to data publishers. Stable and mature software solutions are available, but require maintenance, investment to guarantee up-time, regular upgrading of hardware, etc. Web-based data markets provide a way of outsourcing these tasks "to the cloud". Similarly, developers of data-centric applications often face the problem of finding, accessing and integrating relevant data. Also for this problem, data markets can contribute to the solution.

2. KASABI

Kasabi[2] is a web-based information marketplace (or data market) that is built on *linked data principles*, i.e., all datasets are represented internally as RDF graphs, and each data item is explicitly identified with a URI. This presents a key differentiator to more conventional data markets, in which each dataset is an isolated silo. In contrast, the approach followed by Kasabi allows resources in two or more datasets to refer to each other across different data publishers and domains, thereby adding value to both datasets. The following sections will discuss the different areas of functionality that Kasabi offers to both developers and data publishers.

2.1 The Dashboard

The central hub in Kasabi for each user follows the familiar dashboard metaphor. From here, users can get an overview

[1] `http://www.w3.org/RDF/`
[2] `http://kasabi.com`

WWW2012 Developer Track, April 18-20, 2012, Lyon, France.

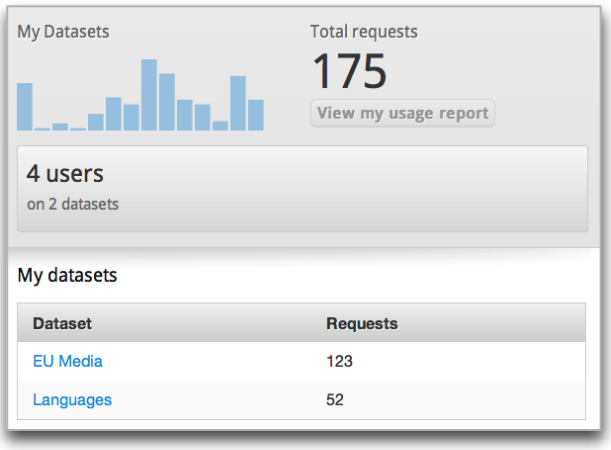

Figure 1: Usage statistics on the dashboard

of their dataset usage (see Fig. 1), create new datasets, update their profile, search for other datasets, etc.

2.2 Publishing Data

When creating a new dataset, Kasabi users provide a set of (mostly optional) metadata, such as a name, textual description, categories, logo and license to apply to the data. Extended documentation for developers using this dataset can also be provided. The actual data is then published to Kasabi as RDF, either through the web interface or the Kasabi API (see below), after which Kasabi will ingest and analyse the dataset. Automatically, a dynamic report card giving a rough overview of the contents is created on the dataset's homepage (see Fig. 2), which also contains all other relevant information about the dataset.

Users can upload arbitrary RDF data to the platform. In this way, it is possible to add existing open data as a Kasabi dataset. E.g., different versions of DBpedia [3] or the Geonames dataset[3] are mirrored in Kasabi. However, Kasabi is able to directly serve each individual resource as linked data, if the resources in the dataset are within the namespace of the dataset in Kasabi. The namespace pattern is as follows:

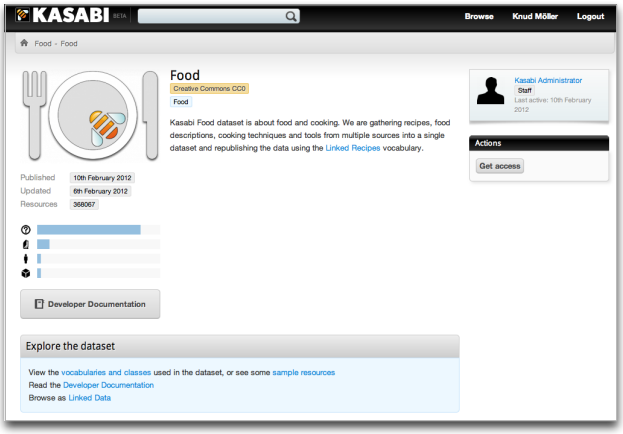

Figure 2: Overview page of the Kasabi Food dataset

```
http://data.kasabi.com/dataset/{dataset_id}/...
```

Following this convention, Kasabi will automatically provide HTML, RDF/XML, RDF/JSON and Turtle representations for each resource in the dataset. Naturally, resources outside the namespace above cannot be served as linked data in this way. However, it is still possible (and for many use cases this might be preferable) to use the various standard APIs provided by Kasabi, as well as to define additional custom APIs over the dataset.

Out of the box, the platform does not perform any kind of inferencing or reasoning over a published dataset. Any kinds of links between resources are considered equal, so that e.g. an `owl:sameAs` link between two resources will be treated the same way as any other type of link. I.e., the two resources are not conflated into one when serving linked data or otherwise exposing data through APIs.

Internally, data in the Kasabi platform is replicated across several nodes for reasons of performance and robustness. This means that, once a dataset has been changed on the platform (data added or removed), these changes need to propagate through to the individual nodes, before the data is consistent throughout the system. In other words, Kasabi uses a so-called "eventual consistency" model for storing data. This means that, in some cases, a successful update may not be immediately visible to all nodes.

2.3 Dataset Descriptions

A common best practice for linked data publishing is the addition of a so-called VoID [2] description, which contains various metadata about the dataset, such as typical example resources, links to APIs and endpoints, licensing information, vocabularies used, etc. Based on a combination of the metadata provided by the user and an analysis of the dataset itself, Kasabi will add such a description and make it available under the dataset's namespace URI (see above) as soon as it is uploaded. E.g., the VoID description for the `world-geography` dataset is available at `http://data.kasabi.com/dataset/world-geography`.

2.4 Dataset APIs

While Kasabi can offer linked data pages only to datasets in its own namespace, it will provide a range of RESTful APIs for any dataset published. Different result formats can usually be specified, such as RDF/XML or JSON. All APIs follow a naming pattern parallel to the dataset naming pattern:

```
http://api.kasabi.com/dataset/{dataset_id}/
   apis/{api_name}
```

A set of five standard APIs is always available:

[3] `http://geonames.org`

```
{
  "head": {
    "pageSize": "10",
    "totalResults": "222",
    "startIndex": "0",
    "query": "apollo"
  },
  "results": [
    {
      "uri": "http://data.kasabi.com/
          dataset/nasa/spacecraft/1968-025A
          ",
      "title": "Apollo 6",
      "score": "1.0"
    },
    {
      "uri": "http://data.kasabi.com/
          dataset/nasa/spacecraft/1975-066A
          ",
      "title": "ASTP-Apollo",
      "score": "0.9938665"
    },
...
}
```


Listing 1: JSON result of Search API

- **Query** — A SPARQL endpoint to the dataset. E.g., `http://api.kasabi.com/dataset/food/apis/sparql`
- **Search** — Keyword-based free-text search against the literals in the dataset. E.g., `http://api.kasabi.com/dataset/food/apis/search` Listing 1 shows an example JSON response to the `nasa` dataset, searching for "apollo".
- **Lookup** — For retrieving a short description of a particular resource in different formats. This is particularly useful for datasets with resources outside the Kasabi namespace, because it offers the equivalent of a linked data page for any arbitrary resource in the dataset. E.g., to retrieve a description of the resource for France (`http://sws.geonames.org/3017382/`) in the Kasabi geonames dataset, one could use the lookup API like this:
  `http://api.kasabi.com/dataset/geonames/apis/lookup?apikey=APIKEY&about=http%3A%2F%2Fsws.geonames.org%2F3017382%2F`
- **Reconciliation** — For resolving literal values such as names or codes against this dataset. Can be used directly in combination with Google Refine[4]. E.g., `http://api.kasabi.com/dataset/food/apis/reconciliation`
  This API facilitates finding links between datasets.
- **Attribution** — A very simple API to attribute the dataset on a webpage by referencing a snippet of Javascript code. E.g., `http://api.kasabi.com/dataset/world-geography/attribution`

In most cases (with the exception of the Attribution API), accessing these interfaces requires the use of an API key, as well as registering with the dataset in question. In addition to the standard APIs, users can also add different kinds of custom contributed APIs, such as stored SPARQL procedures (effectively a short-hand for SPARQL queries) or a Linked Data API (a way to circumvent the restriction of only generating linked data pages for resources local to Kasabi). Figure 3 shows an example of two stored SPARQL procedures for the `world-air-travel` dataset, giving users of the dataset a convenient way to look up lists of airports by their call sign or the country they are located in. Such contributed APIs can be created directly within Kasabi, and follow the same URI pattern as the standard APIs.

[4] `http://code.google.com/p/google-refine/`

Figure 3: Contributed APIs in the World Air Travel dataset


```
{
  "created": "2011-05-27T14:30:00Z",
  "startTime": "2011-05-27T14:31:00Z",
  "endTime": "2011-05-27T14:33:00Z",
  "status": "succeeded"
}
```


Listing 2: JSON response to monitor job status

2.5 Data Management APIs

In addition to the dataset-specific APIs described above, Kasabi also offers functionality through a range of generic data management APIs, such as an update API for the uploading of data or a jobs API for handling various dataset tasks (e.g., resetting a dataset) asynchronously. Whenever a user performs a data management task with these APIs, they will get a URI of the corresponding change resource as a response. E.g., after performing a reset to the `food` dataset, the user might get this URI in response:

`http://data.kasabi.com/dataset/food/jobs/3247389`

Performing a GET request to this resource will then provide the user with a JSON response, indicating whether or not the job has finished, how long it took, etc. An example of this is shown in List. 2. This can be particularly useful for expensive jobs such as dataset updates.

All APIs are RESTful, and can be accessed directly via HTTP, or through one of the available Kasabi client libraries. There are currently client libraries in the Ruby, Javascript, PHP and Python languages. All APIs are documented in detail on the Kasabi website[5].

To illustrate how developers can interact with Kasabi through one of the client libraries, List. 3 gives a brief example of how

[5] `http://kasabi.com/doc/api`

```
require 'rubygems'
require 'kasabi'
require 'json'

#change to your API key
APIKEY = "YOUR_API_KEY"

dataset = Kasabi::Dataset.new
  ("http://kasabi.com/dataset/nasa",
   :apikey => APIKEY )

#Search for "Apollo", limiting to 5 results
results = dataset.search_api_client.search
  ("apollo", { :max => 5 })

#Dump the JSON response to the console
puts JSON.pretty_generate(results)
```

Listing 3: Searching for resources in the apollo dataset

to access the Search API of the `apollo` dataset through the Ruby client library.

2.6 Curated Datasets

To serve as a hub and reference point for data publishers, and to provide curated, cleaned data for various domains to application developers, the Kasabi team regularly adds new vertical datasets. Rather than mirroring a particular data source (e.g., a "geonames" dataset), these datasets cover specific domains in a source-independent way (e.g., the "world-geography" dataset). Where derived from an external source dataset, the curate dataset will keep links to the original resources. At the time of writing, these curated datasets include `food`, `world-air-travel` and `world-geography`. While new datasets are being added to this set, the existing ones are kept up-to-date and consolidated on a regular basis.

3. CONCLUSION

In this short paper, we have given an overview of the Kasabi information marketplace. Kasabi, which is built on linked data principles, offers a wide range of functionality for both data publishers and application developers. Data publishers are relieved from the burden of setting up, maintaining and upgrading a complete hosting environment. Instead, they can do "linked data as a service". At the same time developers are provided with the tools to find, access and integrate data required for building data-rich applications, using a range of both automatically and custom created RESTful APIs.

4. REFERENCES

[1] Digital agenda: Turning government data into gold. European Commmission Press Release, December 2011. `http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1524`.

[2] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing linked datasets with the VoID vocabulary. Interest group note, W3C, March 2011. `http://www.w3.org/TR/void/`.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC+ASWC2007), Busan, South Korea*, pages 11–15. Springer, November 2007.

[4] T. Berners-Lee. Linked data, 2006. `http://www.w3.org/DesignIssues/LinkedData.html`.

[5] S. Harris and A. Seaborne. SPARQL 1.1 query language. Working draft, W3C, January 2012. `http://www.w3.org/TR/sparql11-query/`.

[6] T. Middleton. European commission to adopt open data strategy. Blog Post, November 2011. `http://blog.okfn.org/2011/11/24/european-commission-to-adopt-open-data-strategy/`.

[7] E. Prud'hommeaux and A. Seaborne. SPARQL query language for RDF. Recommendation, W3C, January 2008. `http://www.w3.org/TR/rdf-sparql-query/`.