Apache Any23

Anything to Triples

Michele Mostarda <me@michelemostarda.it>, Twitter: @micmos

Outline

- Scenario
- From Web 1.0 to Web 3.0
- What is RDF
- Machine Readable Annotations
- What is Apache Any23
- Supported Formats
- How does it work
- Web and REST UI
- CLI
- Next Steps
- History of Any23
- How to Contribute
- Slowdown of Web 3.0 and raise of Knowledge Graph

Scenario

Thanks to the promotion of a common set of machine readable annotation standards such as **Microformats** and **RDFa**, sponsored by the main global Web players, we've observed, starting from 2009, a large adoption of embedded markup for data structures (products, reviews, posts, people, organizations, events) in worldwide websites.

The adoption of machine readable markup is the key enabler to the **Web of Data**.

For further details about trends please see reference [1].

From Web 1.0 to Web 3.0

- Web 1.0: Web of interconnected **readonly content**.
- Web 2.0: the **read/write Web**. Pages are scriptable and dynamic, APIs enabling cross interactivity.
- Web 3.0: the **Web of Data**. Complete decoupling data from presentation. Enables the **read/write/understand** mode.

What is RDF (1)

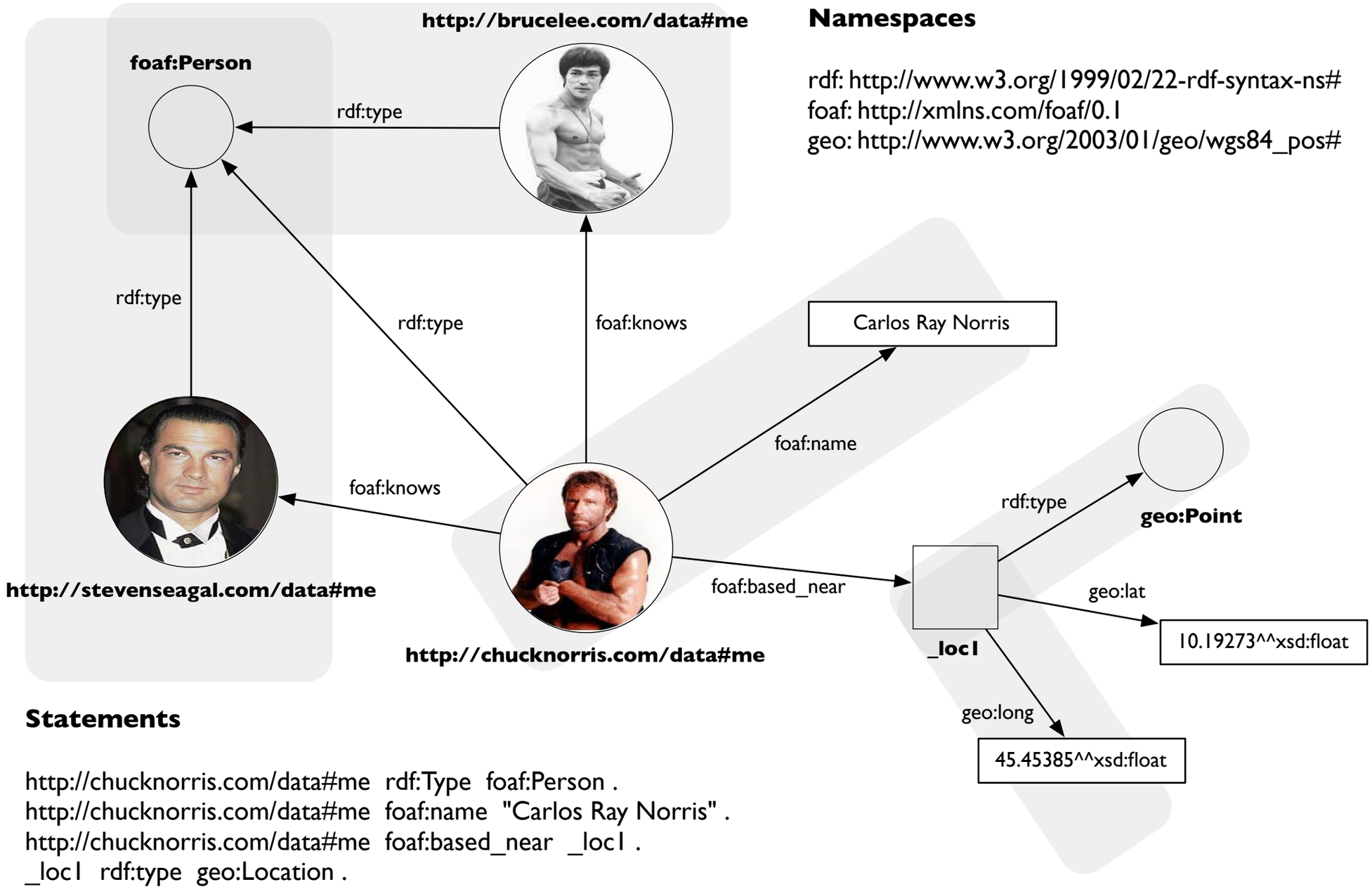
The **Resource Description Framework** is a formalism to describe structured data as a graph.

In this graph nodes are both Web entities (URIs) or literal data (primitive data types) and edges are ontology properties described as URIs.

Every graph is a composition of statements <Subject, Predicate, Object> where *Subject* is both an Entity or a Blank Node, Object can be an Entity, a Blank Node or a Literal.

RDF was adopted as a W3C recommendation in 1999. The RDF 1.0 specification was published in 2004, the RDF 1.1 specification in 2014.

What is RDF (2)



Namespaces

rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
foaf: http://xmlns.com/foaf/0.1
geo: http://www.w3.org/2003/01/geo/wgs84_pos#

Statements

http://chucknorris.com/data#me rdf:Type foaf:Person .
http://chucknorris.com/data#me foaf:name "Carlos Ray Norris" .
http://chucknorris.com/data#me foaf:based_near _loc1 .
_loc1 rdf:type geo:Location .
_loc1 geo:lat 10.19273^^xsd:float .
_loc1 geo:long 45.45385^^xsd:float .
http://chucknorris.com/data#me foaf:knows http://stevenseagal.com/data#me .
http://chucknorris.com/data#me foaf:knows http://brucelee.com/data#me .
http://stevenseagal.com/data#me rdf:Type foaf:Person .
http://brucelee.com/data#me rdf:Type foaf:Person .

Machine Readable Annotations

Enable machines (web crawlers, web scrapers, browsers, etc) to understand page contents in unambiguous way.

Usually expressed as invisible HTML markup around visible content.

Microformats

- First attempt to standardize structured markup in Web content.

- Don’t specifies any ontology, just a simplified data model.

```
<address>
 <a href="http://guptapromoters.com/">Gupta Promoters</a>
</address>
```

```
<address class="vcard">
 <a class="fn url" href="http://tantek.com/">Tantek Çelik</a>
</address>
```

http://microformats.org/

RDFa

- **RDF in Attributes**. Set of standard (W3C) XML attributes to embed RDF statements within visual markup.

- W3C Specification with many variants:
  - RDFa 1.0: applicable only on XML formats (SVG, XHTML)
  - RDFa 1.1: extension of RDFa 1.0 to non XML language domains.
  - RDFa Lite: minimal subset of attributes for general markup operations.

https://www.w3.org/TR/rdfa-syntax/

Microdata

- W3C Specification concurrent to (and simplifying) RDFa.
- Mainly adopted as Open Standard by main search engines with the publication of Schema.org.

https://dev.w3.org/html5/md-LC

Schema.org (1)

Joint effort of Google, Microsoft, Yahoo and Yandex to define a general purpose aligned vocabulary for RDFa, Microdata and JSON-LD.

http://schema.org/

Schema.org (2)

Without Markup | Microdata | RDFa | JSON-LD

```
<div>
  <h1>Fondue for Fun and Fantasy</h1>
  <p>Fantastic and fun for all your cheesy occasions.</p>
  <p>Open: Daily from 11:30am till 11pm</p>
  <p>Phone: 555-0100-3333</p>
  <p>View <a href="http://example.com/menu">our menu</a>.</p>
</div>
```

Without Markup | Microdata | RDFa | JSON-LD

```
<div itemscope itemtype="http://schema.org/Restaurant">
  <h1 itemprop="name">Fondue for Fun and Fantasy</h1>
  <p itemprop="description">Fantastic and fun for all your cheesy occasions.</p>
  <p>Open: <time itemprop="openingHours" datetime="Mo,Tu,We,Th,Fr,Sa,Su 11:30-23:00">Daily from 11:30am till 1
  <p>Phone: <span itemprop="telephone" content="+155501003333">555-0100-3333</span></p>
  <p>View <a itemprop="menu" href="http://example.com/menu">our menu</a>.</p>
</div>
```

Without Markup | Microdata | RDFa | JSON-LD

```
<div vocab="http://schema.org/" typeof="Restaurant">
  <h1 property="name">Fondue for Fun and Fantasy</h1>
  <p property="description">Fantastic and fun for all your cheesy occasions.</p>
  <p>Open: <time property="openingHours" datetime="Mo,Tu,We,Th,Fr,Sa,Su 11:30-23:00">Daily from 11:30am till 1
  <p>Phone: <span property="telephone" content="+155501003333">555-0100-3333</span></p>
  <p>View <a property="menu" href="http://example.com/menu">our menu</a>.</p>
</div>
```

http://schema.org/Restaurant

What is Apache Any23

Any23 is a library, a Web Service and a Command Line Tool written in Java (1.6), that extracts structured RDF data from a variety of Web documents and markup formats.

Any23 is an Apache Software Foundation top level project.

http://any23.apache.org/

Purpose

Purpose of Any23 is manifold, principal applications can be identified in the following areas:

- **Semantic Web scraping**: scraping of of entire websites (or contiguous parts of them) to get a materialization of the underlying Web 3.0 data structures.

- **Page navigation enhancement end enrichment**: identification of relevant entities during page navigation to be used to perform live content enhancement (point out relevant information) and enrichment (add relevant information from other sources).

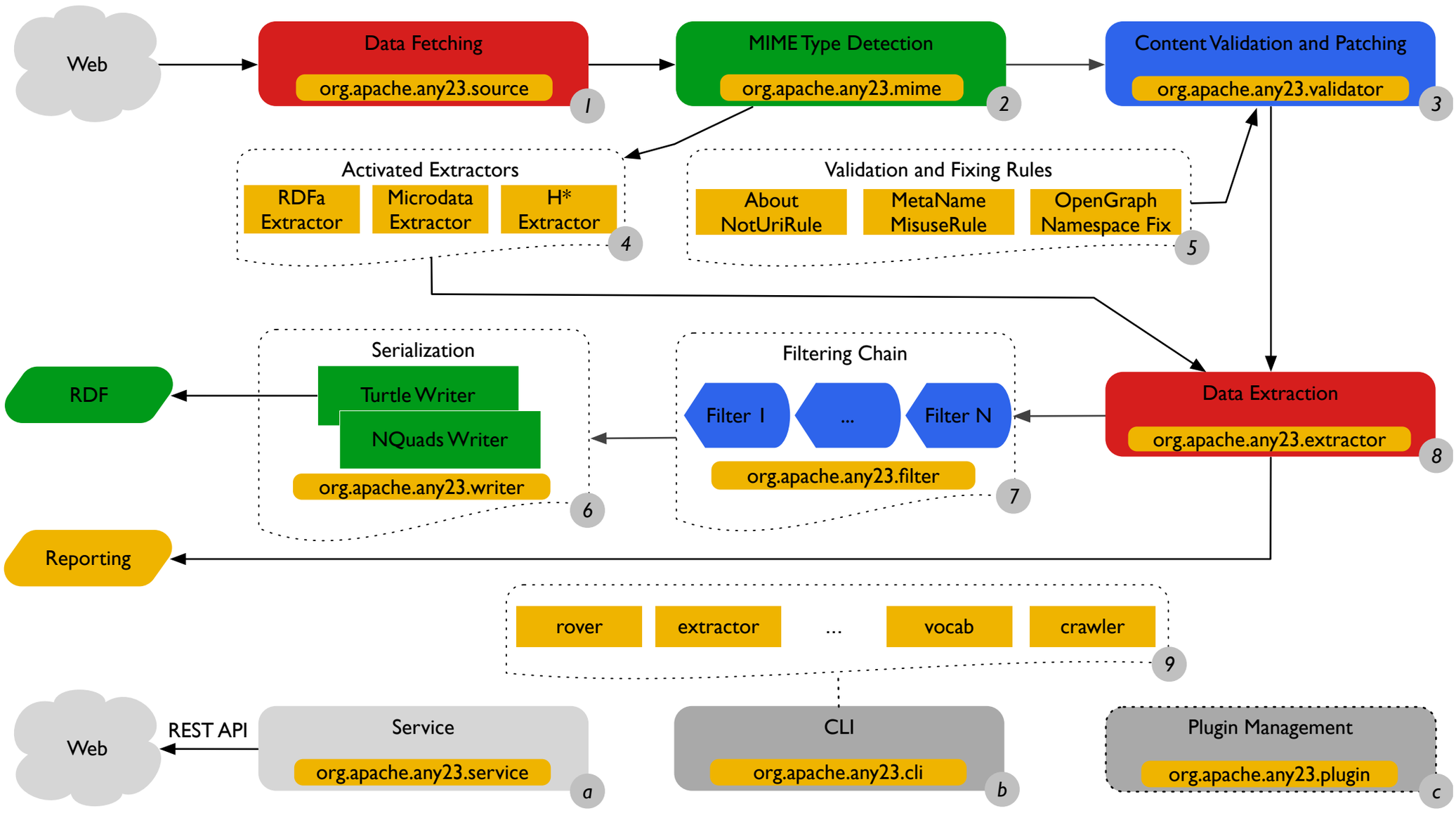
- **Data mashup**: mix of contents from different pages and websites to get a unified intelligent data view.

Supported Formats

Any23 currently supports the following input formats.

- RDF Specific formats: RDF/XML, Turtle, Notation3.
- RDFa with RDFa1.1 (prefix mechanism).
- Microformats: Adr, Geo, hCalendar, hCard, hListing, hRecipe, hReview, License, XFN and Species.
- HTML5 Microdata, such as Schema.org.
- JSON-LD: JSON for Linking Data: a lightweight Linked Data format based on the already successful JSON format and provides a way to help JSON data interoperate at Web-scale.
- CSV: Comma Separated Values.
- Vocabularies: Extraction support for CSV, Dublin Core Terms, Description of a Career, Description Of A Project, Friend Of A Friend, GEO Names, ICAL, Ikif-core, Open Graph Protocol, BBC Programmes Ontology, RDF Review Vocabulary, schema.org, VCard, BBC Wildlife Ontology and XHTML.

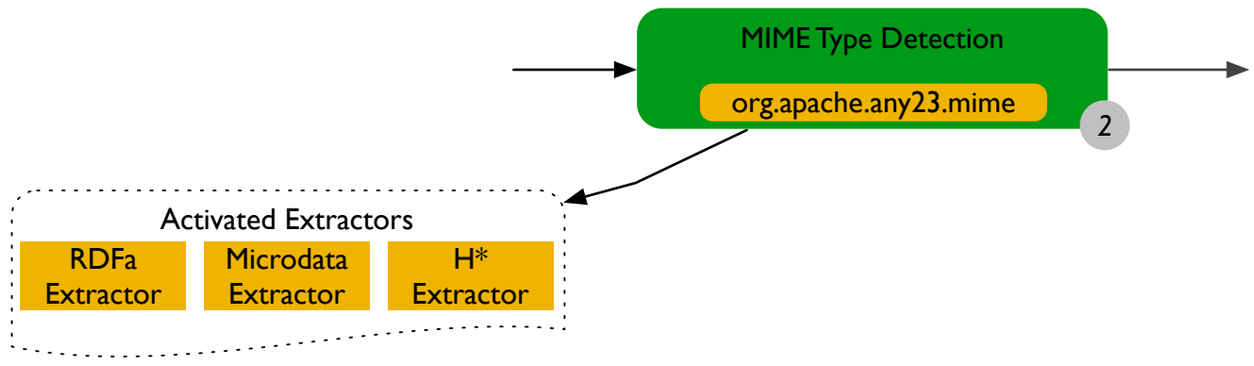
How does it work



Data Fetching

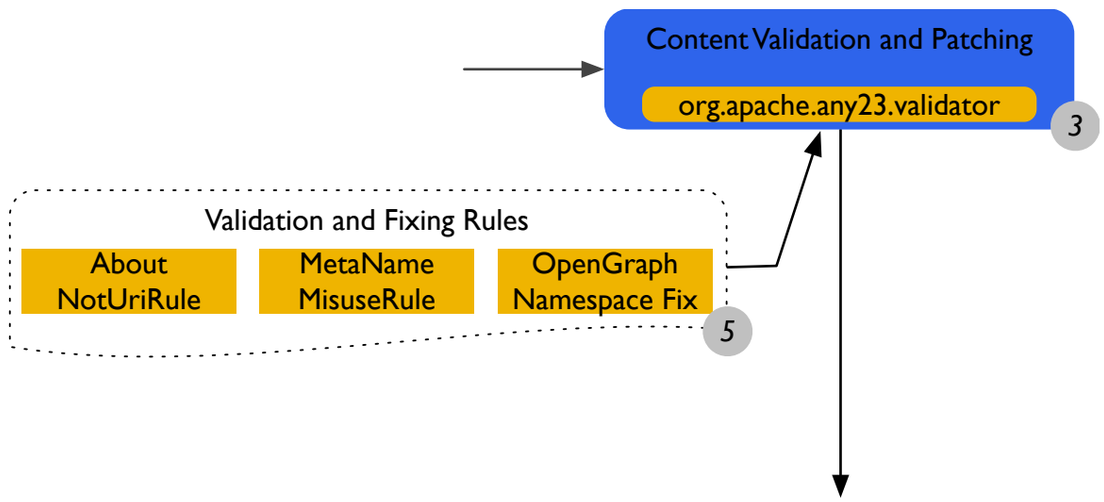
Responsible for negotiating content retrieval with data source.

MIME Type Detection



- Identify data MIME Type if not explicitly specified during fetching phase.
- Uses Apache Tika, a rule-based content analysis toolkit.

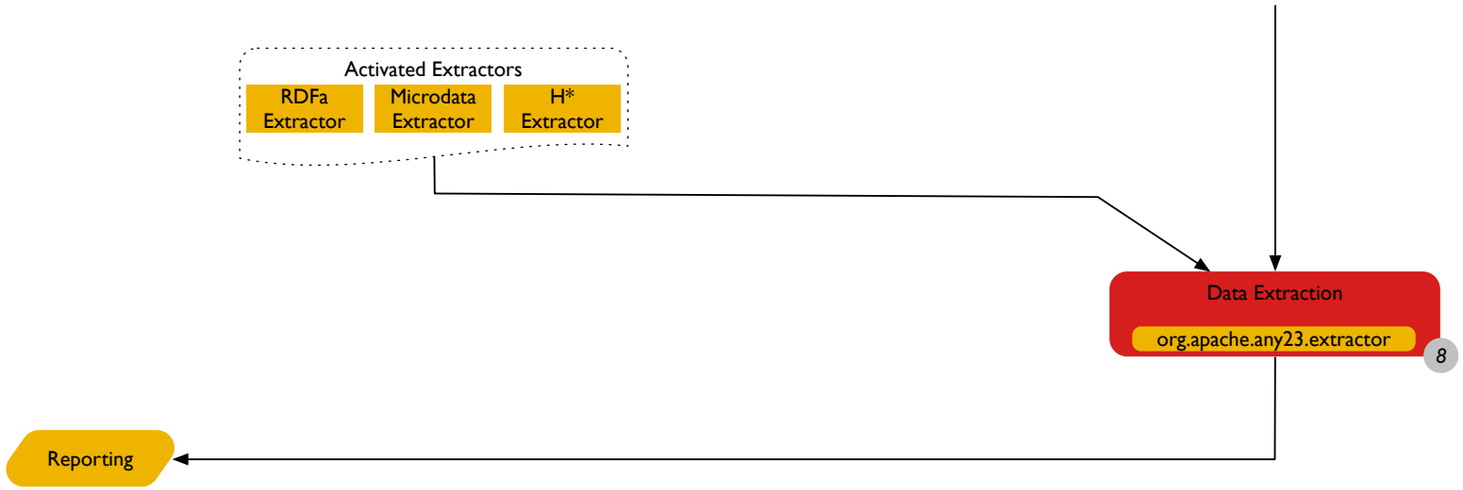
Content Validation and Patching



- Most Web markup contains errors or well known structural issues, time to fix them before parsing.
- Use an extensible rule based engine to perform some common fixes.

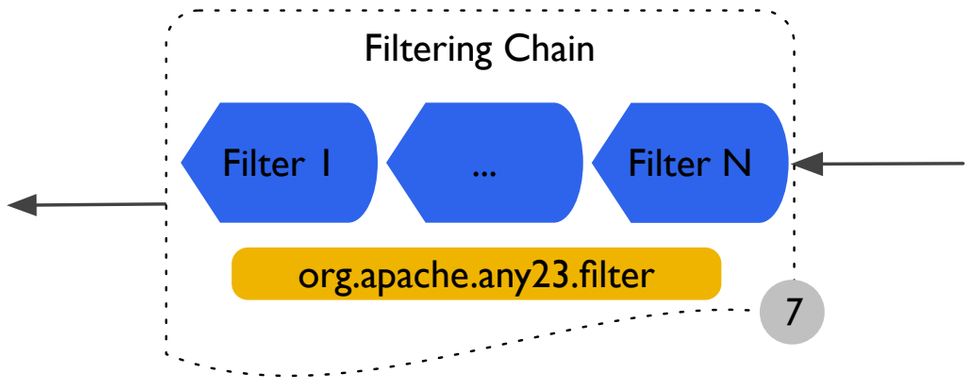
For further details about common web markup issues please refer to reference [3].

Data Extraction



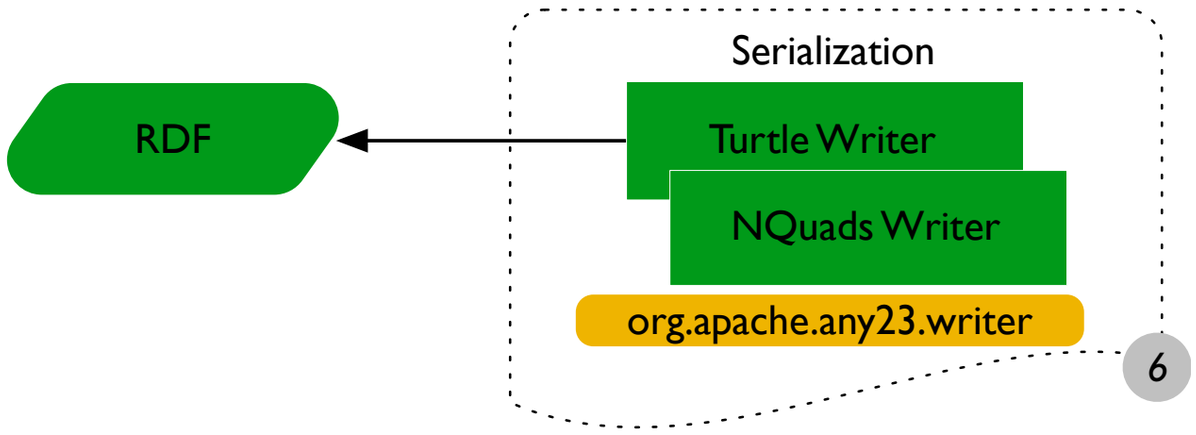
- All Extractors associated to the detected MIME-Type are activated.
- These Extractors inspect the source content, detect specific data structures and emit these as RDF statements.
- Extractors also provide warnings and errors whether they cannot accomplish their tasks.

Filtering



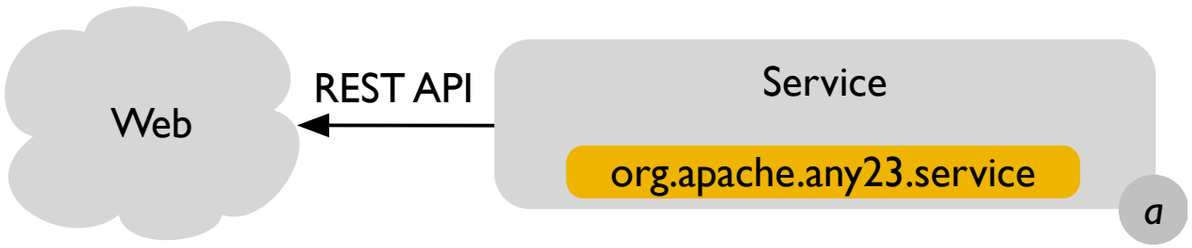
- Optionally it is possible to exclude some tipe of information which is not interesting for the purpose of extraction.
- Filtering is based on RDF Ontology prefixing or on programmatic Filter logic.

Serialization



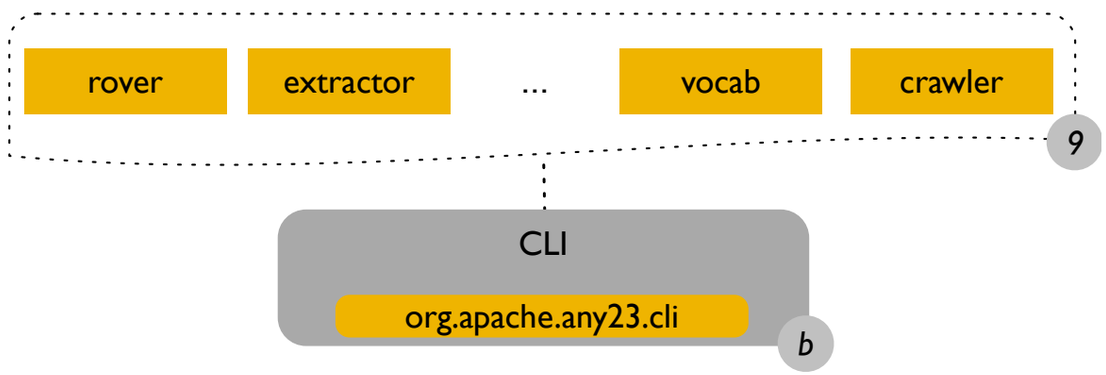
- The statements generated along the pipeline can be finally converted in some of the available RDF transport formats.

Service Module



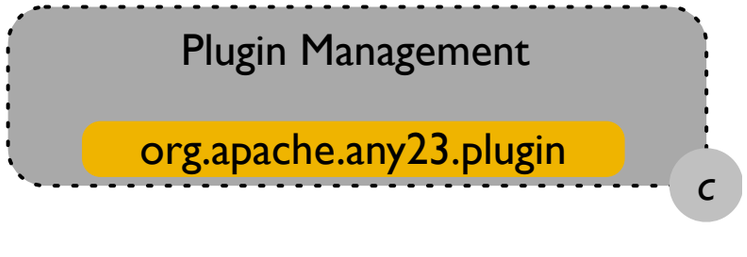
- Provides the REST API and the Web UI.

CLI Module



- Command Line Interface for Any23.
- Exposes main functionalities like:
  - rover: equivalent to Any23 REST Service
  - extractor: supports CLI testing for specific extractor
  - vocab: access to the entire RDF vocabulary supported by Any23 modules
  - crawler: simple Web crawler to scrape markup content from small websites.

Plugin Management



Most Any23 components (Extractors, Filters, ValidationRules, Fixes) can be added as external plugins just registering them in the Any23 classpath.

Web and REST UI

Web UI (1)

Apache Any23 - Anything To Triples - Live Service

Parses Microformats, RDFa, Microdata, RDF/XML, Turtle, N-Triples, JSON-LD and NQuads.

Download and install Any23: visit the Developers Site and the Documentation.

Convert document at URI

Pick an output format and enter the URI of a web document:

http://localhost/ best /

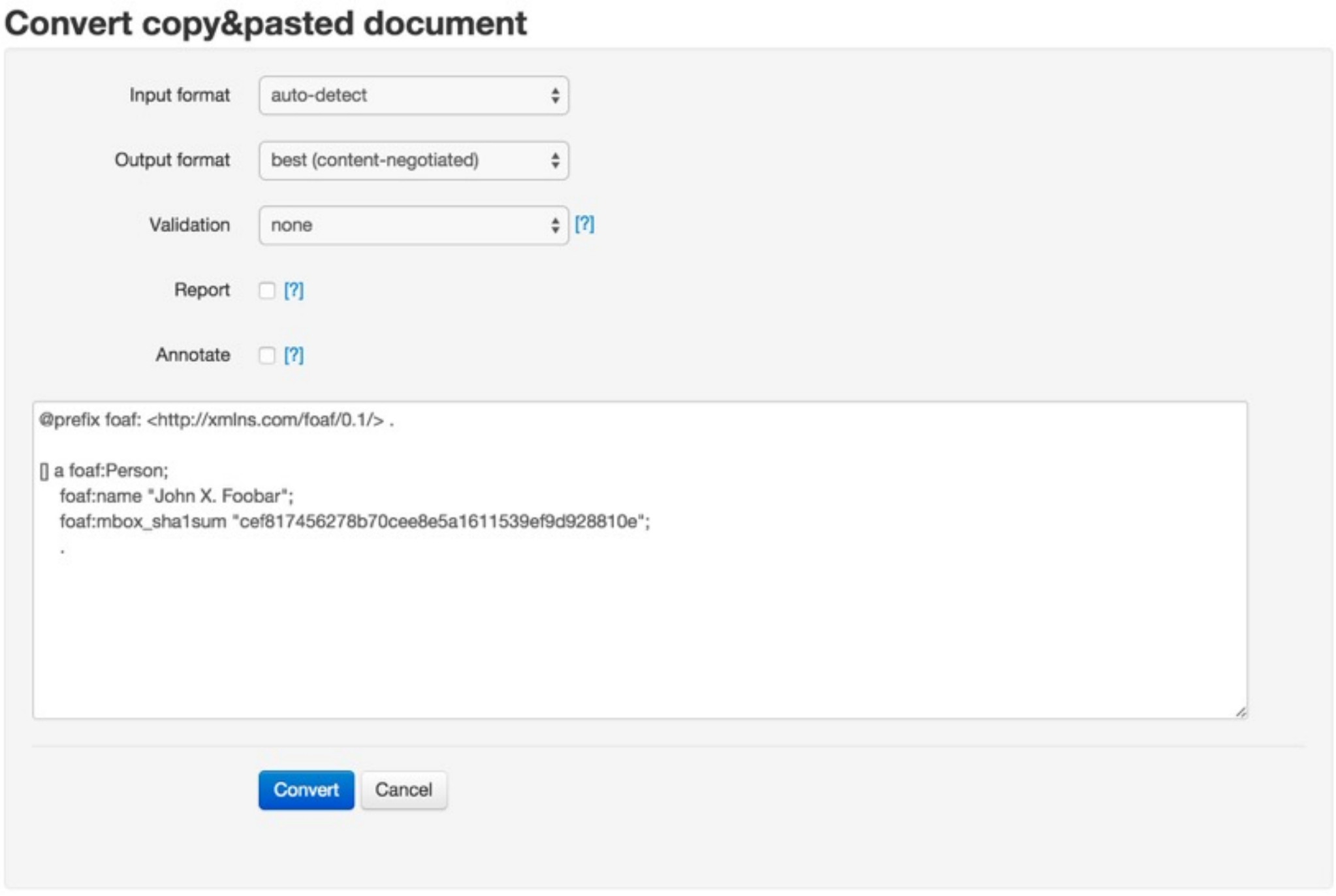
Validation none [?]

Report [?]

Annotate [?]

Convert Cancel

Web UI (2)


Convert copy&pasted document
Input format
auto-detect
Output format
best (content-negotiated)
Validation
none
[?]
Report
[?]
Annotate
[?]
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
[] a foaf:Person;
foaf:name "John X. Foobar";
foaf:mbox_sha1sum "cef817456278b70cee8e5a1611539ef9d928810e";
.
Convert
Cancel

REST UI (1)

Form-style POST API

A document body can also be converted by HTTP POSTing form data to `http://localhost:8080/apache-any23-service/`.
The `Content-Type` HTTP header must be set to `application/x-www-form-urlencoded`. The following parameters are supported:

type	Media type of the input, see the table above. If not present, auto-detection will be attempted.
body	Document body to be converted.
format	Desired output format; defaults to `best`.
validation-mode	The validation level to be applied on the input. Possible values: `none` (no validation applied); `validate` (apply validation and produce validation report if `annotate` flag is enabled); `validate+fix` (apply validation, try to fix detection issues and produce validation report if `annotate` flag is enabled).
annotate	If specified the output RDF will contain extractor specific scope comments. Possible values: `on` / `off`
report	If specified will produce a full XML report containing extraction and validation issues other than produced metadata. Possible values: `on` / `off`

REST UI (2)

Direct POST API

HTTP POSTing a document body to `http://localhost:8080/apache-any23-service/format` will convert the document to the specified output format.
The media type of the input has to be specified in the `Content-Type` HTTP header. Depending on the servlet container, a `Content-Length` header specifying the length of the input document in bytes might also be required.
Typical media types for supported input formats are:

Input format	Media type
HTML	`text/html`
RDF/XML	`application/rdf+xml`
Turtle	`text/turtle`
N-Triples	`text/nt`
N-Quads	`text/nq`
TriX	`application/trix`

Example POST request:

```
POST /apache-any23-service/rdfxml HTTP/1.0
Host: localhost:8080
Content-Type: text/turtle
Content-Length: 174

@prefix foaf: <http://xmlns.com/foaf/0.1/> .

[] a foaf:Person;
    foaf:name "John X. Foobar";
    foaf:mbox_sha1sum "cef817456278b70cee8e5a1611539ef9d928810e";
    .
```

REST UI (3)

Output formats

Supported output format identifiers are:

- `best` for content negotiation according to the client's `Accept` HTTP header
- `turtle`, `ttl`, `n3` for Turtle/N3
- `ntriples`, `nt` for N-Triples
- `nquads`, `nq` for N-Quads
- `trix` for TriX
- `rdfxml`, `rdf`, `xml` for RDF/XML
- `json` for JSON

Error reporting

Processing errors are indicated via HTTP status codes and brief `text/plain` error messages. The following status codes can be returned:

Code	Reason
200 OK	Success
400 Bad Request	Missing or malformed input parameter
404 Not Found	Malformed request URI
406 Not Acceptable	None of the media types specified in the `Accept` header are supported
415 Unsupported Media Type	Document body with unsupported media type was POSTed
501 Not Implemented	Extraction from input was successful, but yielded zero triples
502 Bad Gateway	Input document from a remote server could not be fetched or parsed

Report Format

The XML report format is subjected to changes. The current content is described in section Any23 Service.

Plain Output Format

```
@prefix : <http://xmlns.com/foaf/0.1/> .
@prefix B: <https://www.w3.org/People/Berners-Lee/> .
@prefix Be: <http://www.w3.org/People/Berners-Lee/> .
@prefix blog: <http://dig.csail.mit.edu/breadcrumbs/blog/> .
@prefix card: <https://www.w3.org/People/Berners-Lee/card#> .
@prefix cc: <http://creativecommons.org/ns#> .
@prefix cert: <http://www.w3.org/ns/auth/cert#> .
@prefix con: <http://www.w3.org/2000/10/swap/pim/contact#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix doap: <http://usefulinc.com/ns/doap#> .
[…]

blog:4 dc:title "timbl's blog" ;
      s:seeAlso <http://dig.csail.mit.edu/breadcrumbs/blog/feed/4> ;
      :maker card:i .

<http://dig.csail.mit.edu/data#DIG> :member card:i .

<http://wiki.ontoworld.org/index.php/_IRW2006> dc:title "Identity, Reference and the Web workshop 2006" ;
      con:participant card:i .

<http://www.ecs.soton.ac.uk/~dt2/dlstuff/www2006_data#panel-panelk01> s:label "The Next Wave of the Web (Plenary
Panel)" ;
      con:participant card:i .

<http://www.w3.org/2000/10/swap/data#Cwm> doap:developer card:i .

<http://www.w3.org/2011/Talks/0331-hyderabad-tbl/data#talk> dct:title "Designing the Web for an Open Society" ;
      :maker card:i .

//www4.wiwiss.fu-berlin.de/booksMeshup/books/006251587X> dc:creator card:i ;
      dc:title "Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web" .

<https://timbl.com/timbl/Public/friends.ttl> a :PersonalProfileDocument ;
      cc:license <http://creativecommons.org/licenses/by-nc/3.0/> ;
      dc:title "Tim Berners-Lee's editable FOAF file" ;
      :maker card:i ;
      :primaryTopic card:i .
[…]
```

Report Output Format

```
<?xml version="1.0" encoding="UTF-8" ?>
<response>
  <extractors>
    <extractor>extractor1-name</extractor>
    <extractor>extractor2-name</extractor>
  </extractors>
  <report>
    <message/>
    <error/>
    <issueReport>
    </issueReport>
    <validationReport>
      <errors>
      </errors>
      <issues>
      </issues>
      <ruleActivations>
      </ruleActivations>
    </validationReport>
  </report>
  <data>
<![CDATA[
plain-output-format-here
]]>
</data>
</response>
```

CLI (1)

Get help

```
bin/any23 –help
```

```
Usage: any23 [options] [command] [command options]
  Options:
    -h, --help
       Display help information.
       Default: false
        --plugins-dir
       The Any23 plugins directory.
       Default: /Users/hardest/.any23/plugins
    -X, --verbose
       Produce execution verbose output.
       Default: false
    -v, --version
       Display version information.
       Default: false
  Commands:
    extractor      Utility for obtaining documentation about metadata extractors.
      Usage: extractor [options] Extractor name
        Options:
          -a, --all
             shows a report about all available extractors
             Default: false
          -i, --input
             shows example input for the given extractor
             Default: false
          -l, --list
             shows the names of all available extractors
             Default: false
[…]
```

CLI (2)

Get input output samples for every registered Extractor

```
bin/anv23 extractor -a
[…]
Extractor: rdf-jsonld
      type: ContentExtractor
-------- Example Input  --------
{
  "@context": {
    "name": "http://xmlns.com/foaf/0.1/name",
    "knows": "http://xmlns.com/foaf/0.1/knows"
  },
  "@id": "http://me.markus-lanthaler.com/",
  "name": "Markus Lanthaler",
  "knows": [
    {
      "@id": "http://manu.sporny.org/about#manu",
      "name": "Manu Sporny"
    },
    {
      "name": "Dave Longley"
    }
  ]
}
[…]
-------- Example Output --------
<http://me.markus-lanthaler.com/> <http://xmlns.com/foaf/0.1/knows> <http://manu.sporny.org/about#manu> ,
_:b0 ;
      <http://xmlns.com/foaf/0.1/name> "Markus Lanthaler"^^<http://www.w3.org/2001/XMLSchema#string> .

<http://manu.sporny.org/about#manu> <http://xmlns.com/foaf/0.1/name> "Manu Sporny"^^<http://www.w3.org/2001/
XMLSchema#string> .

_:b0 <http://xmlns.com/foaf/0.1/name> "Dave Longley"^^<http://www.w3.org/2001/XMLSchema#string> .

[…]
```

CLI (3)

Extract all possible data from a specific URL

```
bin/any23 rover http://www.w3.org/People/Berners-Lee/card
```


```
{
      "quads": [
            [{
                  "type": "uri",
                  "value": "http://www.w3.org/DesignIssues/Overview.html"
            }, "http://purl.org/dc/elements/1.1/title", {
                  "type": "literal",
                  "value": "Design Issues for the World Wide Web",
                  "lang": null,
                  "datatype": null
            }, null],
            [{
                  "type": "uri",
                  "value": "http://www.w3.org/DesignIssues/Overview.html"
            }, "http://xmlns.com/foaf/0.1/maker", {
                  "type": "uri",
                  "value": "https://www.w3.org/People/Berners-Lee/card#i"
            }, null],
            [{
                  "type": "uri",
                  "value": "http://www.w3.org/People/Berners-Lee/card"
            }, "http://www.w3.org/1999/02/22-rdf-syntax-ns#type", {
                  "type": "uri",
                  "value": "http://xmlns.com/foaf/0.1/PersonalProfileDocument"
            }, null],
[…]
```

CLI (4)

Dump the Any23 internal vocabulary (in Turtle format)

```
bin/anv23 vocab -f turtle
<http://www.estrellaproject.org/lkif-core/role.owl#Role> a <http://www.w3.org/
2000/01/rdf-schema#Class> ;
    <http://www.w3.org/2000/01/rdf-schema#member> <http://www.estrellaproject.org/
lkif-core/role.owl#> .
<http://www.estrellaproject.org/lkif-core/role.owl#Function> a <http://www.w3.org/
2000/01/rdf-schema#Class> ;
    <http://www.w3.org/2000/01/rdf-schema#member> <http://www.estrellaproject.org/
lkif-core/role.owl#> .
<http://www.estrellaproject.org/lkif-core/role.owl#Social_Role> a <http://www.w3.org/
2000/01/rdf-schema#Class> ;
    <http://www.w3.org/2000/01/rdf-schema#member> <http://www.estrellaproject.org/
lkif-core/role.owl#> .
[…]
```

Next Steps

- Introduce a scriptable DSL to speedup development of Extractors, Rules and Fixes.
- Introduce a public repo to croudify plugin development leveraging on cloning / pulling mechanism.

How to Contribute

http://any23.apache.org/developers.html

http://www.apache.org/foundation/how-it-works.html

Bit of history

- Any23 starts in 2008 in DERI (now Insights Centre, insight-centre.org), become part of the sindice.com data acquisition pipeline.

- In 2010 is promoted as Apache Incubator project (incubator.apache.org/projects/any23.html), in 2012 is it promoted to top level project (any23.apache.org).

Sindice.com and SindiceTech

- **Sindice.com** has been an attempt to build the first Semantic Search Engine. Operating from 2007 to 2014 under Insight Centre (former DERI), at the end it had indexed 700M semantically annotated Web pages with a refresh rate of 20M pages per day for a total of 6B statements.

- **SindiceTech** is a startup with mission of commercialize the technologies developed along the Sindice.com experience.

Slowdown of Web 3.0 and Raise of Knowledge Graph

- Many issues in massive adoption of Machine Readable embedded markup in Web contents.
- Most data publishers **don’t have the culture of structured data**.
- Even worst, most data publishers add markup only for a small subset of data available in their Web pages because **afraid of the loss of traffic**.
- the Web of Data model **still needs to define a valuable business model**.
- Since 2012, thanks to the Schema.org initiative and HTTP friendly formats like JSON-LD, the Web of Data is living a new momentum.
- Schema.org and other collective initiatives like WikiData are the pillars of the **Google Knowledge Graph**, which is a proprietary alternative to DBpedia and Linked Data.

References

- [1] Web Data Commons - RDFa, Microdata, and Microformat Data Sets (http://webdatacommons.org/structureddata)

- [2] Deployment of RDFa, Microdata, and Microformats on the Web – A Quantitative Analysis (http://hannes.muehleisen.org/Bizer-etal-DeploymentRDFaMicrodataMicroformats-ISWC-InUse-2013.pdf)

- [3] Heuristics for Fixing Common Errors in Deployed schema.org Microdata (http://dws.informatik.uni-mannheim.de/fileadmin/lehrstuehle/ki/pub/MeuselPaulheim-HeuristicsForFixingCommonErrorsInDeployedSchemaOrgMicrodata-ESWC2015.pdf)