International Semantic Web Conference

Thursday, October 29, 2009

Discovering and Maintaining Links on the Web of Data

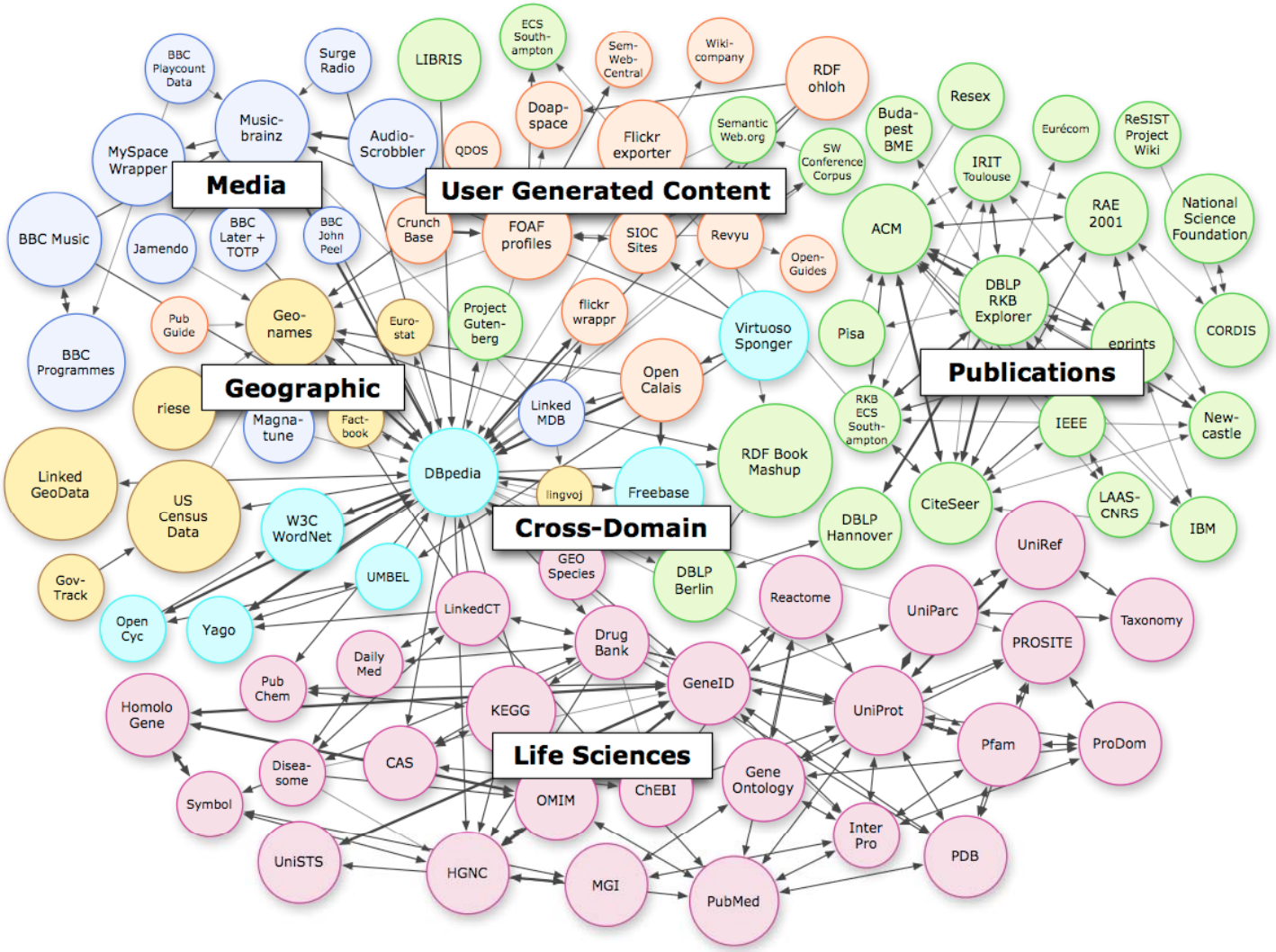
Julius Volz, Chemnitz University of Technology

Christian Bizer, Freie Universität Berlin

Martin Gaedke, Chemnitz University of Technology

Georgi Kobilarov, Freie Universität Berlin

The Semantic Web is a single global data space because data sources are connected by LINKS



Two bad things about Links

1. **Setting links is effort for data publishers**
   - **Currently handled using custom code**
2. **Links have to be updated as data sources change**
   - **Currently not handled at all :-(**

Outline

1. **Silk – Link Discovery Framework**
2. **Web of Data – Link Maintanance Protocol**

The Silk Link Discovery Framework

- Open source tool for discovering relationships between entities within different Linked Data sources.
- Makes it easier for data publishers to set RDF links from their data sources to other data sources on the Web.

- Main Features
  1. Can generate owl:sameAs and other link types
  2. Flexible, declarative language for specifying link conditions
  3. works with local and remote SPARQL endpoints
  4. works in situations where terms from different schemata are mixed
  5. uses local caching, indexing and pre-matching to increase performance

Link Conditions

- specify which conditions two entities must fulfill in order to be interlinked.
- are expressed as a combination of
  - RDF path expressions
  - similarity metrics and
  - aggregation functions.

Supported Similarity Metrics

`qGramSimilarity`	string similarity based on q-grams
`jaroSimilarity`	string similarity based on the Jaro distance metric
`jaroWinklerSimilarity`	string similarity based on the Jaro-Winkler metric
`stringEquality`	returns 1 if strings are equal, 0 otherwise
`numSimilarity`	percentual numeric similarity
`dateSimilarity`	similarity of date values
`uriEquality`	returns 1 if URIs are equal, 0 otherwise
`taxonomicSimilarity`	concept similarity based on taxonomic distance
`maxSimilarityInSet`	highest similarity encountered by comparing a single item to all items in a set
`setSimilarity`	similarity between two sets of items

Supported Aggregation Functions

- **Similarity values may be aggregated using:**
  - AVG – weighted average of similarity value set
  - MAX – choose highest similarity value in set
  - MIN – choose lowest similarity value in set
  - EUCLID – Euclidian distance aggregation
  - PRODUCT – weighted product of similarity value set

- **Weights & optional metrics**
  - metrics may be weighted in AVG, EUCLID and PRODUCT aggregations
  - metrics may be
    - declared optional
    - given a default value

    in case a metric fails to evaluate because of missing RDF values

Path-based Selector Language

- for addressing the graph around an entity
- Operators

/	Forward step
\	Backward step
[ ]	Filter

- Examples

```
# Select the labels of the directors of a movie
?movie/dbpedia:director/rdfs:label

# Select the albums of a given artist
?artist\dbpedia:artist[rdf:type = dbpedia:Album]
```

Supported Transformation Functions

`removeBlanks`	Remove whitespace from string
`removeSpecialChars`	Remove special characters from string
`lowerCase`	Convert a string to lower case
`upperCase`	Convert a string to upper case
`stem`	Apply word stemming to a string
`alphaReduce`	Strip all non-alphabetic characters from a string
`numReduce`	Strip all non-numeric characters from a string
`replace`	Replace all occurrences of a string
`translateWithDictionary`	Translate string using a dictionary provided as comma separated values

Example: Linking DBpedia and GeoNames

- **Data sources**
  1. DBpedia – data extracted from Wikipedia (2.6 million concepts)
  2. GeoNames – open geographical database (6.5 million locations)
- **Objective**
  - Discover links between cities in DBpedia and GeoNames!
- **In this example, we want to**
  - discover `owl:sameAs` links
  - compare
    - city names,
    - geo coordinates,
    - population counts
    - wikipedia links
  - use thresholds

Silk – Linking Specification

```
<?xml version="1.0" encoding="utf-8" ?>
<Silk>
    <Prefix id="rdf"
     namespace="http://www.w3.org/1999/02/22-rdf-syntax-ns#" />
    ...

    <DataSource id="dbpedia">
        <EndpointURI>http://dbpedia.org/sparql</EndpointURI>
        <Graph>http://dbpedia.org</Graph>
        <DoCache>1</DoCache>
        <PageSize>1000</PageSize>
    </DataSource>

    <DataSource id="geonames">
        <EndpointURI>http://geonames.org/sparql</EndpointURI>
    </DataSource>
    ...
```

Define Namespaces

Specify SPARQL endpoints

Silk – Linking Specification

```
...
<LinkSpec id="cities">
    <LinkType>owl:sameAs</LinkType>

    <SourceDataset dataSource="dbpedia" var="a">
        <RestrictTo>
            { ?a rdf:type dbpedia:City }
            UNION
            { ?a rdf:type dbpedia:PopulatedPlace }
        </RestrictTo>
    </SourceDataset>

    <TargetDataset dataSource="geonames" var="b">
        <RestrictTo>
            ?b gn:featureClass gn:P
        </RestrictTo>
    </TargetDataset>
...
```

Specify link type

Specify source dataset

Specify target dataset

Silk – Linking Specification

```
...
<LinkCondition>
    <AVG>
        <MAX>
            <Compare metric="jaroSimilarity" optional="1">
                <Param name="str1" path="?a/rdfs:label[@lang 'en']" />
                <Param name="str2" path="?b/gn:alternateName[@lang 'en']" />
            </Compare>
            <Compare metric="jaroSimilarity" optional="1">
                <Param name="str1" path="?a/rdfs:label" />
                <Param name="str2" path="?b/gn:name" />
            </Compare>
        </MAX>
        <MAX>
            <Compare metric="numSimilarity" optional="1">
                <Param name="num1" path="?a/dbpedia:populationEstimate" />
                <Param name="num2" path="?b/gn:population" />
            </Compare>
            <Compare metric="numSimilarity" optional="1">
                <Param name="num1" path="?a/dbpedia:populationTotal" />
                <Param name="num2" path="?b/gn:population" />
            </Compare>
        </MAX>
```

Compare city names Using Jaro Similarity

Aggregate results

Compare populations

Silk – Linking Specification

```
            ...
            <Compare metric="maxSimilarityInSets" optional="1" weight="1">
                <Param name="set1" path="?a/foaf:page" />
                <Param name="set2" path="?b/gn:wikipediaArticle" />
                <Param name="submetric" value="stringEquality" />
            </Compare>
            <Compare metric="numSimilarity" optional="1" weight="0.3">
                <Param name="num1" path="?a/wgs84_pos:lat" />
                <Param name="num2" path="?b/wgs84_pos:lat" />
            </Compare>
            <Compare metric="numSimilarity" optional="1" weight="0.3">
                <Param name="num1" path="?a/wgs84_pos:long" />
                <Param name="num2" path="?b/wgs84_pos:long" />
            </Compare>
        </AVG>
    </LinkCondition>
    <Thresholds accept="0.9" verify="0.7" />
    <Limit max="1" />
    <Output acceptedLinks="accepted_links.n3" format="n3"
            verifyLinks="verify_links.n3" />
  </Interlink>
</Silk>
```

Compare links to Wikipedia

Weight results

Compare geo-coordinates

Specify thresholds, link limits and output format

Results of the Experiment

Compared RDF resources

- 40,197 cities in DBpedia
- 2,410,855 populated places in Geonames

Generated links

- 35,012 links above the *accept* threshold of 0.9
- 3,246 links between the *accept* and and *verify* (0.7) thresholds

Examples of discovered links

```
<http://dbpedia.org/resource/Berlin> owl:sameAs
<http://sws.geonames.org/2950159/> .
```

```
<http://dbpedia.org/resource/Chemnitz> owl:sameAs
<http://sws.geonames.org/2940132/> .
```

Evaluate the Quality of Generated Links

Silk:web | Load Config | Compare Resources | Evaluate Links

Compare Resource Pairs

Link Triples (.nt):

```
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00814> <http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Meloxicam> .
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01036> <http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Tolterodine> .
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00378> <http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Dydrogesterone> .
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00948> <http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Fulvestrant> .
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01033> <http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Metoclopramide> .
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00025> <http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Anakinra> .
```

LinkSpec ID: drugs ☐ Reverse pair order Compare!

High: 0.999999839972 **Low:** 0.224517813463 **Average:** 0.626288195802 **Exceptions:** 0

Source Resource	Target Resource	Similarity (click row for details)
http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00378	http://dbpedia.org/resource/Dydrogesterone	0.9999998399723472
http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01036	http://dbpedia.org/resource/Tolterodine	0.9999996927690115
http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00814	http://dbpedia.org/resource/Meloxicam	0.9999985771322385
http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00948	http://dbpedia.org/resource/Fulvestrant	0.2728216702567859
http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00025	http://dbpedia.org/resource/Anakinra	0.26039158122051537
http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01033	http://dbpedia.org/resource/Metoclopramide	0.22451781346294356

/resource/drugs/DB00947	http://dbpedia.org/resource/Fulvestrant	0.9999995879836248
http://www4.wiwiss.fu-berlin.de/drugbank /resource/drugs/DB00814	http://dbpedia.org/resource/Meloxicam	0.9999985771322385
http://www4.wiwiss.fu-berlin.de/drugbank /resource/drugs/DB00025	http://dbpedia.org/resource/Anakinra	0.26039158122051537

```
AVG(
  MAX(
    maxSimilarityInSets(
      submetric = "jaroSets",
      set1 = path("?a/rdfs:label"),
      set2 = path("?b/rdfs:label")
    ) = 0.650793650794 [W:1.0 O:False D:None],
    maxSimilarityInSets(
      submetric = "jaroSets",
      set1 = path("?a/rdfs:label"),
      set2 = path("?b/drugbank:synonym")
    ) = 0.5375 [W:1.0 O:True D:None],
    maxSimilarityInSets(
      submetric = "jaroSets",
      set1 = path("?a/rdfs:label"),
      set2 = path("?b/drugbank:genericName")
    ) = 0.650793650794 [W:1.0 O:True D:None]
  ) = 0.650793650794 [W:1.0 O:False D:None],
  stringEquality(
    str2 = path("?b/drugbank:pubchemCompoundId")
  ) = None [W:5.0 O:True D:None],
  numSimilarity(
    num1 = path("?a/dbpedia-owl:molecularweight"),
    num2 = path("?b/drugbank:molecularWeightAverage")
  ) = 0.0651905464339 [W:2.0 O:True D:None]
) = 0.260391581221 [W:- O:- D:-]
```

http://www4.wiwiss.fu-berlin.de/drugbank /resource/drugs/DB01232	http://dbpedia.org/resource/Metoclopramide	0.1625713291831919

Compare Link Sets

Silk:web | Load Config | Compare Resources | Evaluate Links

Evaluate Links

Correct Links (.nt): (clear)

```
<http://dbpedia.org/resource/Anakinra> .
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00357>
<http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Aminoglutethimide> .
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00639>
<http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Butoconazole> .
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01320>
<http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Fosphenytoin> .
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01505>
<http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Etoxeridine> .
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01157>
<http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Gemifloxacin> .
```

Generated Links (.nt): (clear)

```
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01233>
<http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Metoclopramide> .
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00026>
<http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Anakinra> .
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00357>
<http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Aminoglutethimide> .
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB00639>
<http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Butoconazole> .
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01320>
<http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Fosphnytoin> .
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01505>
```

☐ Reverse pair order Evaluate! **Precision:** 0.75 **Recall:** 0.8181818181818182 **F1-measure:** 0.7826086956521738

Missing Links (2 of 11):

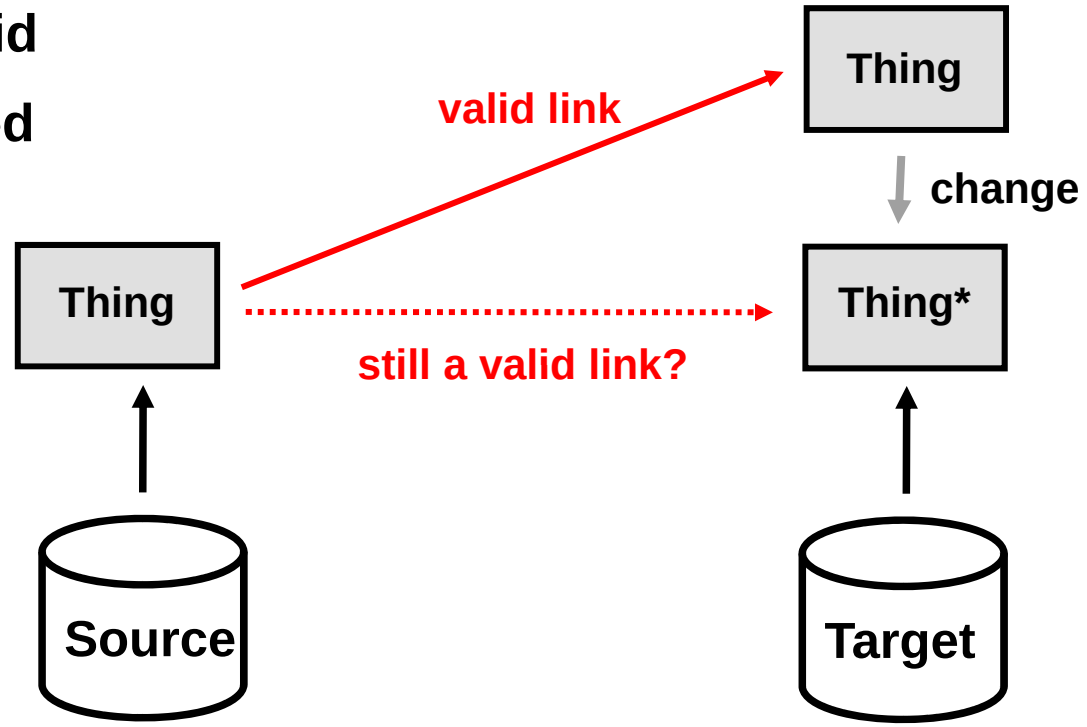
```
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01320>
<http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Fosphenytoin>
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01157>
<http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Gemifloxacin>
```

Incorrect Links (3 of 12):

```
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01320>
<http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Fosphnytoin>
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB01155>
<http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Gemifloxacin>
<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/DB02901>
<http://dbpedia.org/about/html/http://www.w3.org/2002/07/owl#sameAs>
<http://dbpedia.org/resource/Dihydrotestosterone>
```

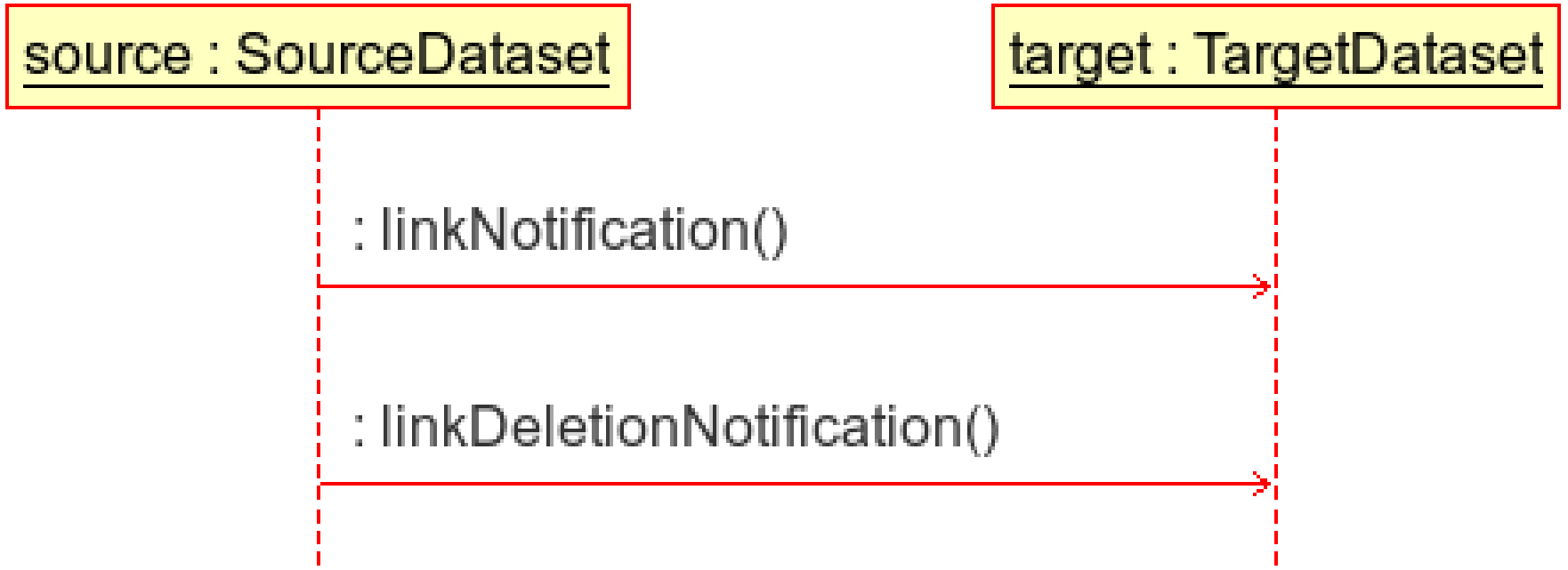
2. WoD - Link Maintenance Protocol

- **Problem: As datasets change over time,**
  1. existing links become invalid
  2. new links may be discovered



- **Solution: Web of Data – Link Maintenance Protocol (WoD-LMP)**
  - automates the communication between cooperationg data sources
  - two roles: link source and link target dataset
  - uses SOAP to exchange messages

1. Notify target about links pointing at it

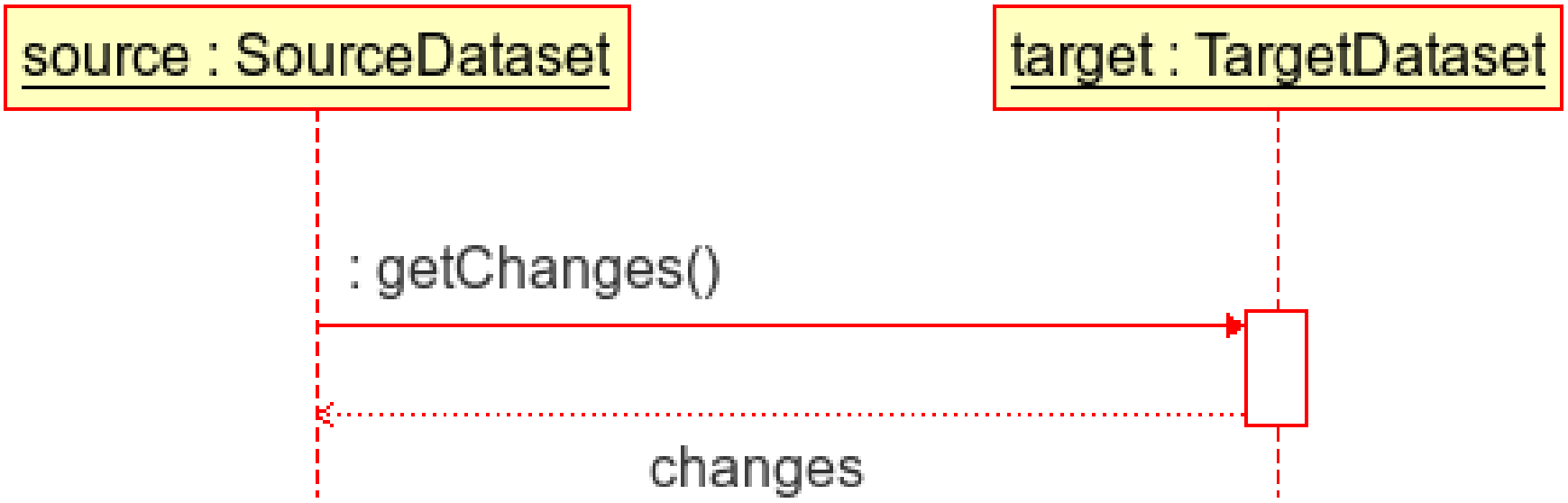


- **Allows target to decide whether it wants to set back links.**

LinkNotification Message

```
<Message type="link_notification">
        <Endpoint uri="source endpoint URI">
        <Link>
            <SourceResource uri="source URI" />
            <TargetResource uri="target URI" />
            <LinkType uri="link type" />
        </Link>
        <Link>
            <SourceResource uri="source URI" />
            <TargetResource uri="target URI" />
            <LinkType uri="link type" />
        </Link>
         ...
 </Message>
```

2. Request list of changes since a point in time



- **Request all changes or specific change types:**
  - **additions, updates and deletions of resources.**
- **The link source may use this change information for periodical link recomputation.**

GetChanges Message

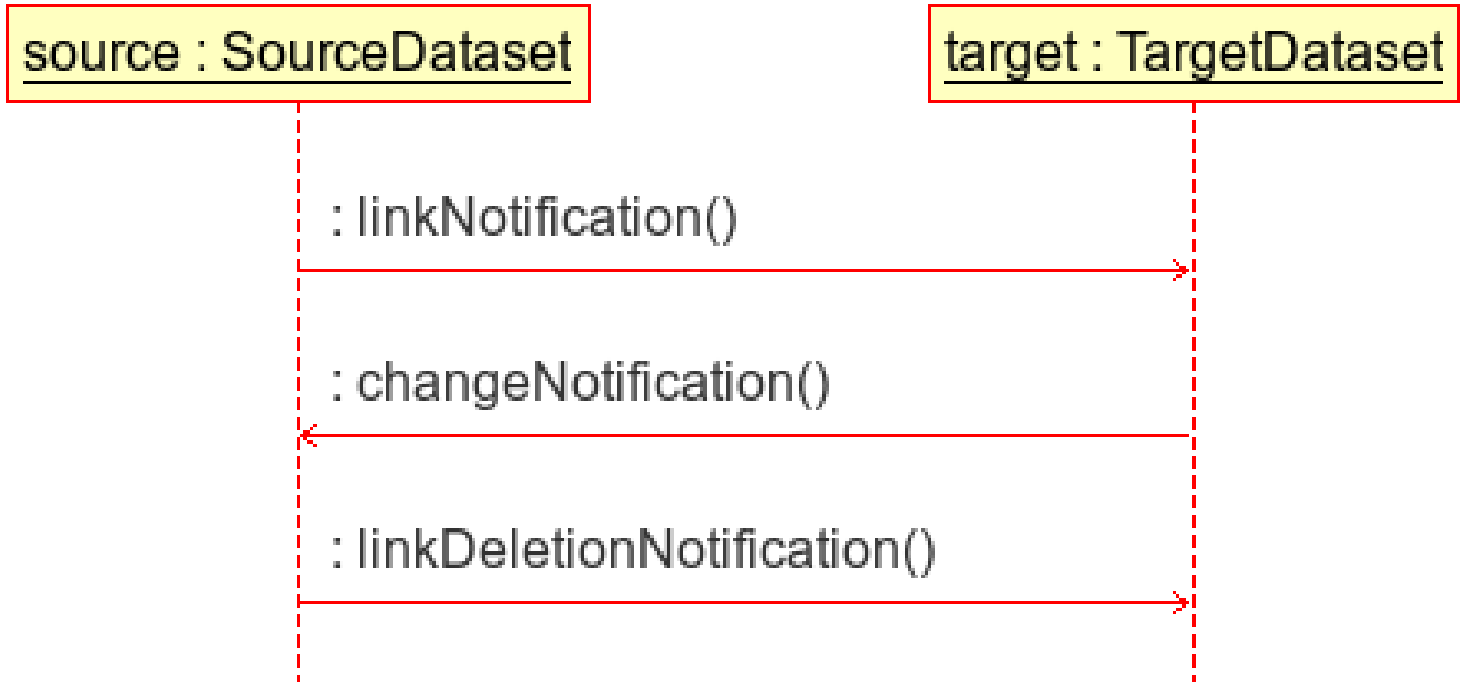
<Message type="change_notification">

<Resource uri="**resource URI 1**" type="**change type**"
seqnum="**seqnum 1**" />

<Resource uri="**resource URI 2**" type="**change type**"
seqnum="**seqnum 2**" />

</Message>

3. Subscribe to Target Changes



- The source subscribes to be informed about changes to resources that were used to compute links.
- The target monitors these resources and notifies the source about changes.

LinkNotification Message

```
<Message type="link_notification">
        <Endpoint uri="source endpoint URI">
        <Link>
            <SourceResource uri="source URI" />
            <TargetResource uri="target URI" />
            <LinkType uri="link type" />
            <SubscribeTo uri="subscribed resource 1" />
            <SubscribeTo uri="subscribed resource 2" />
        </Link>
        <Link>
          ...
       </Link>
 </Message>
```

Outlook

- **Implement more metrics and aggregation functions**
  - **lots of existing work in databases and ontology matching**
- **Further improve performance**
  - **Lots of existing work in databases**
- **Use link maintenance protocol to**
  - **Maintain links between Linking Open Drug Data (LODD) data sources**
  - **Maintain links between DBpedia and other data sources**

Thanks!

References

- Silk - Language Specification
  http://www4.wiwiss.fu-berlin.de/bizer/silk
- Web of Data – Link Maintanance Protocol Specification
  http://www4.wiwiss.fu-berlin.de/bizer/silk/wodlmp/
- Download Silk and WoD-LMP (BSD License)
  http://silk.googlecode.com/
- Linked Data Overview Article
  Bizer, Heath, Berners-Lee: Linked Data – The Story So Far
  http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf