DELIVERABLE D4.4

Tailored search application on top of data published using ODN – final

<table>
<thead>
<tr>
<th>Project</th>
<th>Components Supporting the Open Data Exploitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acronym</td>
<td>COMSODE</td>
</tr>
<tr>
<td>Contract Number</td>
<td>FP7-ICT-611358</td>
</tr>
<tr>
<td>Start date of the project</td>
<td>1st October 2013</td>
</tr>
<tr>
<td>Duration</td>
<td>24 months, until 31st September 2015</td>
</tr>
<tr>
<td>Date of preparation</td>
<td>22.6.2015</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Michiel Hildebrand, Roberto Cornacchia</td>
</tr>
<tr>
<td>Responsible of the deliverable</td>
<td>Roberto Cornacchia</td>
</tr>
<tr>
<td>Email</td>
<td><a href="mailto:roberto@spinque.com">roberto@spinque.com</a></td>
</tr>
<tr>
<td>Reviewed by</td>
<td>Jan Kučera, Ivan Hanzlik</td>
</tr>
<tr>
<td>Status of the Document</td>
<td>final version</td>
</tr>
<tr>
<td>Version</td>
<td>1.2</td>
</tr>
<tr>
<td>Dissemination level</td>
<td>PU (Public)</td>
</tr>
</tbody>
</table>
Table of Contents

1 Executive summary
2 Introduction
3 Developments on the Spinque software stack
   3.1 Spinque 2.0
      User Interface
      Multiple output streams
      Multiple items per stream
      Improved hierarchical reasoning
      SKOS support
      More power to SpinQL
   3.2 Features in testing phase
4 Linked Open Images
   4.1 Frontend: mobile application
   4.2 Backend: Spinque strategies
5 LINK
   5.1 Existing alignment tools
   5.2 User Interface
   5.3 A Cultural Heritage Example
6 Summary
1 Executive summary

Deliverable D4.4 is about one end of the open data supply-demand chain addressed in COMSODE: applications for end-users. The objective of this deliverable is described in the DOW as: Tailored search application on-top of data published using ODN – final: Freely accessible search application based on SPINQUE’s “Search by Strategy” concept, tailored to open data published as part of the COMSODE project. Its output includes prototype applications that can show the feasibility of the Open Data Node approach, from the point of view of application development.

After the proof-of-concept deliverable D4.2 of the first year of the project, this deliverable includes the following achievements:

- New technology developed, which will be part of the exploitation strategy
- Validation of ODN output (D4.3)
- Validation of the Search by Strategy approach in the context of Open Data Node. The same technology has been used to develop both final search applications and support tools for ODN (vocabulary alignment)
- Support of COMSODE pilot cases

The applications developed (described below) are available online:

1. Linked Open Images: http://link.spinque.com/openbeelden
2. CultuurLINK: http://cultuurlink.beeldengeluid.nl
2 Introduction

In the first year of the project we adapted the Spinque software to make it compatible with the Open Data Node, and Linked Data in general. We demonstrated the technical feasibility of the approach with the development of two prototype applications, described in COMSODE deliverable D4.2\textsuperscript{1}. The restaurant search application already covered the full pipeline to create advanced search applications on top of ODN with Spinque. We used various data sources from the Open Data Node, we indexed the data in Spinque, we created the tailored API required by the application by modeling search strategies, and developed a simple user interface.

The aim of Spinque is that any developer or information specialist can use the Spinque software platform to create the tailored API they need for an application or service. For example, when an institution needs a new service on top of the data in ODN they create the required functionality by modeling strategies, following the Search by Strategy\textsuperscript{2} approach. A third party developer that is creating a mobile App on top of ODN creates the entire backend required by the App by modeling strategies in Spinque. In the second year of the project we managed to get Spinque closer to this goal.

We extended the Spinque core software stack to better support the process of modeling strategies over Linked Data. Using the strategy editor the user had to be familiar with the structure and content of the data when modeling strategies. When dealing with heterogeneous Linked Data the user might, however, not be familiar with all data sources, making it difficult to determine what the strategy should look like and how to model it. Therefore, we developed a first prototype of the strategy editor with a new user interface. We adopted a data driven approach that helps the user inspect data sources and select directly from the data the relations and values that are needed in the strategy. We also improved the indexing and reasoning support for typical Linked Data constructs. Making it easier to model complex strategies, for example including sub-property reasoning or hierarchies in SKOS vocabularies.

We demonstrate the improved Spinque software with the development of a mobile application on top of the ODN. In a pilot with the user board member the Netherlands Institute for Sound and Vision we collected several Dutch cultural heritage datasets, we created Linked Data out of them, and provided integrated access through a mobile application. With the application titled Linked Open Images the user explores historic material about World War II. It gives access to historic Dutch News reels. When watching the video of a news reel the user can, at any time, pull up a panel to explore related information. This exploratory mode extends the experience with immediate access to the most related books and photographs. In addition, the application suggests other news reels about the same topic, for the user to watch next.

Within the second year of the project we also developed an application to support data providers with data integration. During the development of the search applications in the first and second

\textsuperscript{2} http://www.spinque.com/concept
year of the project we were required to integrate the datasets used in the applications. We found the existing tool support too limited for our needs. In the first year of the project we ended up creating scripts to create links between data sources. In several commercial projects we have similar experiences. Data integration is required and we often end up creating ad hoc solutions. Therefore, we decided to develop an application to support the linking process using Spinque’s Search by Strategy approach. The result is a new product we call LINK.

In the remainder of this deliverable we describe the development on the Spinque software, the Linked Open Images application and LINK.
3 Developments on the Spinque software stack

During the first year of COMSODE, Spinque developed a version of their search platform that is specifically designed to address RDF and Linked Data, in order to ensure a smooth interaction with Open Data Node. This prototype platform, internally named Spinque LD (Linked Data), introduced a number of changes to the Spinque Core platform, which we can summarise as follows:

- data connectors for Linked Data sources (RDF formats, SPARQL end-points);
- extension to the internal data-model to support the standardised RDF triple model;
- fully automatic indexing for RDF data;
- initial support for hierarchical reasoning by efficiently indexing rdfs:subPropertyOf and rdfs:subClassOf properties;
- building blocks for specific RDF constructs.

From this baseline, the development roadmap has been driven by two factors:

- the improvements necessary to support the applications described in this document;
- turning the initial prototype for Linked Data into a mature technology ready for exploitation.

Goodbye Spinque LD, welcome Spinque 2.0. During the second year of development, we realised that efficient support for Linked Data should become part of our main platform. Therefore, the experience collected and the improvements achieved during the development of the internal project Spinque LD are being merged into our core system. This is resulting in a new major release of Spinque Core, named Spinque 2.0.

Spinque 2.0 is currently being used by Spinque for all new projects, including the applications described in this document. It will be released officially by the end of COMSODE.

3.1 Spinque 2.0

User Interface

A new prototype for a User Interface has been designed and as a first step applied only to the CultuurLINK application (described later in this document).

Besides more modern visual components and a cleaner interface, the new UI aims at simplifying the usage of the Strategy Editor, by bringing the following improvements:
- the data / result preview has a more central role, for a more intuitive overall process;
- intermediate result inspection is more effective thanks to real-time interaction features;
- the novel concept of building strategies starting from simple interactions with data has been introduced.

The rationale driving the development new interface was to ease the task to develop search strategies by third parties.

**Multiple output streams**

Each building block can now output any number of result streams of any primitive type, which can be labelled.

In the simple example on the right, block Has property filters incoming items based on whether or not they have a skos:altLabel property defined. The block outputs both the items that satisfy the required match (RESULT) and the ones that do not (NOT).

This new feature immediately improves the experience of editing search strategies, providing more expressive power and making strategies easier to read.

**Multiple items per stream**

The previous extension has been taken one step further in Spinque 2.0, by supporting multiple items per stream. The example on the left shows two incoming streams, A and B. The goal is to identify matching pairs, based on the value of their properties, respectively rdfs:label and skos:prefLabel. The three outputs, NOTA, NOTB, RESULT, describe respectively the items from A that do not match, those from B that do not match, and the pairs of items that do match.

The colour-coded pins show clearly that each pair in the output stream RESULT contains elements from both A and B.

In all building blocks, any pin, both in input and in output, can be defined as a tuple (a pair in this example) of items. This extension allows to design strategies where building blocks can transport not only results from one another, but also contextual information.

**Improved hierarchical reasoning**

Basic support for hierarchical reasoning was introduced in the first version of Spinque LD. In Spinque 2.0, this support has been consolidated and made available to all main building blocks.

The example on the right shows the configuration of a block to match items from two separate data streams, A and B. The match is based on the values of property skos:prefLabel for stream A, and rdfs:label for stream B. However, the USE_SUBPROPERTIES flag indicates to also consider the value of all sub-properties on both sides.
This allows to write effective strategies also without detailed knowledge of how data was modeled. Any property `mySchema:myLabel` that is defined as a sub-property of `rdfs:label` would be captured by the configuration in the example.

SKOS support

SKOS (Simple Knowledge Organization System) is used to describe relationships among concepts expressed by any type of structured controlled vocabulary. Explicit support for SKOS has been introduced in Spinque 2.0:
- with specific building blocks able to navigate the hierarchical structures defined by `skos:broader` and `skos:narrower` properties;
- with functionalities to export new match definitions, based on the `skos` properties in the screenshot.

More power to SpinQL

SpinQL is the Domain Specific Language that is used inside building blocks to program their inputs, outputs, and actions. It borrows concepts from Probabilistic Relational Algebra and translates into SQL database queries. The language processor has undertaken a major rewrite, in order to offer better support for the new features listed above and others that are still in development.

Some of the main improvements are:
- support for macro definitions and conditional constructs, to better support building block parameter configuration
- transparent intermediate result caching mechanism across sessions and strategies
- improved translation speed

3.2 Features in testing phase

The features listed below are being tested at the time of this deliverable and are expected to be included in the official release of Spinque 2.0.

- On-demand indexing
  Some search functionalities rely on indexing structures being present at query-time. An example is keyword search that relies on text statistics. The advanced result caching mechanism already implemented (see above), allows to specify building blocks that don't rely on any pre-computed statistics. These will be computed transparently when

---

3 A probabilistic relational algebra for the integration of IR and DB systems, Fuhr and Rölleke, 1997
needed and stored for later re-use. On-demand indexing enables full flexibility (no need to know which information to pre-compute), more elegant and declarative building block specification, and potentially more efficiency (compute only what is needed)

- **More granularity in text search**
  Currently, one has to decide at indexing time which data fragments can be used for keyword search and what is the language of the relative text. The version under development allows to search any fragment of text in any language, also based on query-time decisions.

- **Simplification of building block library**
  After a heavy rewrite of the building block library, this is being simplified and cleaned-up, to make it more usable by customers and partners.

- **New User Interface**
  The new Strategy Editor UI used in CultuurLINK application has passed the experimental phase and is expected to become the standard one in the next release.
4 Linked Open Images

Linked Open Images is a mobile application to explore historic material about World War II. It gives access to three Open Data collections. First, a collection of historic Dutch news reels shared by the Netherlands Institute for Sound and Vision. Second, a collection of Photographs from the NIOD Institute for War, Holocaust and Genocide Studies. Third a collection of books from the same NIOD institute. The collections are already provided as Open Data by the institutes through the OAI-PMH harvesting protocol. This data is, however, less rich than originally maintained by the institutes. For example, metadata terms from the institute’s controlled vocabularies are not provided in the open data versions. In addition, each institute uses their own controlled vocabularies, named GTAA and the NIOD term list respectively. To provide integrated access to these collections we first used the Open Data Node to create Linked Open Data out of the collections and vocabularies. Next, we indexed the data with Spinque and created the tailored API required by the application by modeling Spinque strategies.

The front-end builds upon CWI’s enriched video experiences, an application developed in LinkedTV (a recently completed EU research project). The user watching a video can, at any time, pull up a panel to explore related information. Within COMSODE we extended the application to operate on mobile phones, tailored the design to the use case and integrated it with Spinque. The exploratory mode in our use case extends the experience of viewing Sound and Vision’s historic news reels with immediate access to the most related NIOD books and photographs. In addition, the application suggests other news reels about the same topic, for the user to watch next.

The entire backend of the application is created by modeling Spinque strategies. In this case, we modeled different strategies to (1) search for videos, (2) find related photographs and books, and (3) suggest related videos. In other words we tailored the generic Linked Video application to this use case by modeling the search strategies. To create a similar user experience for different collections the only prerequisite is to model the search; a novel workflow we call search by strategy!

4.1 Frontend: mobile application

The application is available at http://link.spinque.com/openbeelden (even though developed for mobile devices, it will work on a desktop or laptop computer as well).
A list of news reels is shown upon opening the application. The user can also search for a specific video with the search bar. The screenshot above shows how the user searches for videos of the liberation of Amsterdam (“bevrijding” in Dutch).
When the user selects a video, the application switches to the player. The screenshot above shows the video player with the video about the liberation of Amsterdam. The video player provides the default functionality (pausing/playing, timebar with seeking functionality, volume control) and in addition contains a button to enable the “explore panel” (situated top right).
This “explore panel” contains three tabs to group different types of related information. The first tab contains related photographs retrieved the NIOD collection, that the user may swipe through. The screenshot above, for example, shows a photograph of a crowd gathering at the Dam Square in Amsterdam right after the liberation. By clicking the info button on one of the photographs, additional information about the photograph is revealed.

Similarly, the second tab contains references to books from the NIOD library. The third and final tab contains related videos, retrieved from the news reel collection. In the screenshot above, the first suggestion is a video of the liberation speech given by the mayor of Amsterdam. Selecting one of the suggested items from this tab loads this video into the player. Of course, while viewing this new video, the user may again open the explore panel and bring up related information to this footage, and so on.

### 4.2 Backend: Spinque strategies

The creation of the application’s backend required no programming whatsoever! All this search-based functionality is readily provided by modelling six different Spinque strategies. The start screen is backed by two search strategies, a first to get the list of initial videos (news reels) and another one to search for videos. The video player uses a third strategy to fetch detailed information about a video, such as the stream URL and the duration of the video. The explore panel is supported by three more strategies that express recommendations, which identify the related information shown in each tab, photographs, books and videos.

Let us elaborate on these three recommendation strategies, that follow the recommendation pattern: as input they take the current item (the video), information is extracted from this item,
and related items are retrieved using this extracted information. Within this pattern, the three strategies differ from each other as the data sources each contain different types of information.

http://www.openbeelden.nl/120155

rdf:type dcmtype:MovingImage
dc:title “Bevrijdingsbeelden Amsterdam”
dc:description “Reportage van de intocht der Canadezen en Britten in de hoofdstad...”
dc:date “8-1-1945”
dc:coverage gtaa:34040, gtaa:31586, ...
dc:subject gtaa:215996, gtaa:217248, ...
dc:publisher: "Nederlands Instituut voor Beeld en Geluid"

Snippet of news reel metadata from the Netherlands Institute for Sound and Vision

http://www.beeldbankwo2.nl/detail_no.jsp?action=detail&imid=110408

rdf:type dcmtype:Image
dc:description “Een menigte mensen op Dam, met name in georganiseerd verband ...”
dc:coverage gtaa:34040, gtaa:31586, ...
dc:subject niod:Bevrijding, niod:Massabijeenkomsten, ...

Snippet of photograph metadata from the NIOD

The strategy to recommend photographs from the NIOD beeldbank given a news reel uses three types of metadata: textual, subject terms and locations. The figures above show a snippet of news reel metadata from the Netherlands Institute for Sound and Vision and one of photograph metadata from the NIOD. For the videos, textual metadata is available in the form of a title (dc:title) and a description (dc:description). The subject terms (dc:subject) and the depicted location (dc:coverage) are taken from the audio visual thesaurus of the institute (GTAA). Subject terms include “liberation”, “parades”, “jeeps”, “street footage”, and “resistance fighters” while locations include “Amsterdam” and “Dam square”. For the photographs the textual
metadata is only available in the form of a description while locations (dc:coverage) are taken from GTAA. For this example, the locations also include “Amsterdam” and “Dam square”. The subject terms (dc:subject) are taken from the NIOD term list, examples including “liberation” and “mass parades”. In the process of making Linked Open Data the terms from the NIOD were linked to the concepts from GTAA.

**Spinque search strategy to get photographs related to a news reel**

The figure above shows the entire strategy for the recommendation of photographs given a video. The strategy starts at the top with a block that holds all data (genericIndex). The next block selects from this data the input video (filter_id_equal). This block is configured as an input block. This means that in the API corresponding to this strategy the value of this block has to be provided as a parameter. In this case the input is provided by the identifier of the video (a URL).

Below the input block the strategy contains the three streams to use the different types of metadata. The blocks for each stream are shown in a different color. The blocks in yellow together find the photographs from which the titles and descriptions are similar to those of the input video. This is achieved by first extracting the terms contained in the title and description of the input video (find_TERM_from_ANY). These terms are then used as a query to find other items. This is done with the block rank_text_BM25. The items that we search for are restricted to photographs by filtering the dataset to items of type Image, using the block filterByType. The output of this stream consists of photographs described by terms similar to the input video. The
related photographs are ranked using the BM25 algorithm. Roughly speaking this algorithm prioritizes the role of the important terms in the recommendation.

In the textual stream (yellow blocks) we had to extract terms from the title and descriptions. In the case of subjects (blue blocks) and locations (green blocks), the terms to use have been modeled explicitly as relations to concepts from the GTAA thesaurus or NIOD term list (dc:subject and dc:location). To find photographs related to a video, we simply traverse the relations over these concepts. For the locations, this is straightforward because videos and photographs are both described using GTAA; we traverse the dc:coverage relation from the video to the GTAA concept (e.g. Dam Square) using the block traverseRelation, and we traverse the same relation in the opposite direction to find all other items related to this Concept traverseRelationBackwards. For the subject terms we need one additional block to traverse the link relation between the concept from GTAA and the NIOD term list. This block is configured to traverse the link between the concepts in the form of the skos:exactMatch relation.

The Linked Open Images application demonstrates how to build advanced search applications over heterogeneous collections without programming. The API powering the Linked Open Images application is created by modeling the search strategies in Spinque. Another important part that was required to develop the application was the integration of the different data sets, linking the controlled vocabularies. In the next section we describe Spinque LINK, the application we developed within COMSODE to support this vocabulary alignment process.
5 LINK

LINK supports data owners with the alignment of their data sources. The need for this application was provided by the COMSODE methodology. An important part of the methodology is to make datasets available in an integrated fashion, D5.4 Methodologies for the integration of datasets. One aspect of this integration is the alignment of datasets by linking entities from one dataset with the same entities in another datasets. Within COMSODE we used LINK to align the different data sources used in the restaurant inspection application and the Linked Culture application.

Together with the Netherlands Institute for Sound and Vision we are deploying LINK into a national service for the Dutch Culture Heritage community. The service, titled CultuurLINK, supports collection owners with the alignment of their vocabularies. It allows them to link their internal vocabularies to several large vocabularies that are core for the community, such as the audiovisual thesaurus provided by the Netherlands Institute for Sound and Vision. In this way institutions can enrich the descriptions of their collections with for example additional background information and labels in multiple languages.

We briefly discuss two existing alignment tools and their limitations that motivated the requirements for LINK, before we describe the user interface of the application and illustrate the functionality with an example from CultuurLINK.

5.1 Existing alignment tools

While several tools exist to perform the alignment of datasets we experienced that they are difficult to apply in a specific use case. In the first year of COMSODE we ended up writing our own ad hoc solutions to find the same restaurant across different databases. Existing tools provide little support to assess the quality of produced results, and it is unclear how the tool should be (re-)configured to improve the results. SILK partially overcomes these limitations as it allows the user to manually construct the workflow to find links⁴. With the graphical strategy editor of SILK the user constructs this workflow out of building blocks. SILK provides powerful building blocks, such as various algorithms to compare strings. However it is not straightforward to construct a workflow for a specific case. The intention of the research prototype Amalgame is to provide an interactive approach to vocabulary alignment⁵. With Amalgame the user also constructs a workflow, as in SILK, but in this case the user can see the intermediate results at each step. This allows the user to quickly try out different strategies. The downside is that Amalgame does not provide all the powerful blocks as SILK.

LINK combines the support for an interactive alignment process as provided by Amalgame, with powerful match techniques found in SILK.

5.2 User Interface

With LINK, the user builds a link strategy step-by-step out of basic building blocks; in a similar fashion as constructing a strategy for a Spinque search engine. The interface is shown in the

⁴ http://silk-framework.com/
⁵ http://semanticweb.cs.wu.nl/amalgame/
The top part of the application contains the strategy editor and the bottom part shows the result table. The editor contains the building block library on the left side and the canvas to the right. The user builds the link strategy step-by-step by dragging building blocks on to the canvas and connect them with the other blocks. CultuurLINK contains blocks to filter data sources, match the items from those by comparing their attributes, filter matches using structural properties and partition the result sets for analysis.

When the user selects a building block, the output is computed and presented in the result table. Inspecting the intermediate results helps the user to decide which step to take next, enabling an interactive approach to vocabulary alignment. When the link strategy is completed, the links are exported as SKOS triples, which can be added to the Open Data Node. The definition of the link strategy provides the provenance of the links and can be exported as well.

5.3 A Cultural Heritage Example

The Linked Open Images application described in Section 4 used the collections of two Dutch institutions, the audiovisual collection of the Netherlands Institute for Sound and Vision and the photograph and library collection of the Netherlands Institute for War, Holocaust and Genocide Studies (NIOD). The collections are described with subject terms from controlled vocabularies. Each institute uses their own controlled vocabularies, named GTAA and the NIOD term list respectively. We used LINK to find the corresponding concepts among these two vocabularies. The screencast at http://www.youtube.com/watch?v=55Wvo1-DZpY. Here we discuss the main steps.

The GTAA and NIOD vocabularies have been modeled using the RDF vocabulary for Simple Knowledge Organization Systems (SKOS). The data sources are represented by the two blocks at the top of the strategy (see also the Figure above). By selecting a datasource block, the...
content is shown in the table at the bottom of the UI. In the Figure, the GTAA is selected and the table shows its SKOS Concepts, 177,987 in total. Columns show the typical SKOS attributes of the concepts, such as the preferred and alternative labels. Below the column header, a chart indicates how many of the concepts have a value for this attribute. This helps the user to understand how useful these attributes are when used to find links.

By inspecting the result table the user also finds out that the GTAA vocabulary contains different types of concepts organized in different concept schemes, e.g. person names, geographical names, subject terms etc. The NIOD term list only contains subject terms. Therefore we first filter the GTAA to subject terms, as matching also on the other types of concepts could introduce noise in the results. The filter is added by dragging a filter block into the strategy canvas, connecting to the datasource block and configuring appropriately.

Blocks are connected by dragging a line from the output connector of one block (shown at the bottom of a block) to the input connect of another block (at the top of a block). The configuration panel of a block is opened by the clicking the config icon at the top right of the block. The figure below shows the configuration of the string match block that we discuss next.

![String match configuration panel](image)

*CultuurLINK: configuration panel of the string match block.*

The next step is to match the concepts from the two vocabularies by comparing their attributes. The user can try different types of string match techniques and apply them to different types of attributes. At each step the instant feedback on the results allows the user to determine if the operation is suited, the configuration should improved, the strategy should be extended, or the block should be discarded in favor of a different strategy.
We first find the concepts that have the exact same preferred labels, by adding a string match block and configuring it match on preferred labels. This results in 557 links for 1343 of the concepts of the NIOD term list. By manually evaluating a small subset it becomes clear these matches can be accepted as good links. The figure below shows the result table with the links found by the string match block. It shows for each link at first row the concept from the NIOD vocabulary and at the second row the matching concept from the GTAA vocabulary. The links are manually evaluated by selecting a relation from the dropdown menu.

<table>
<thead>
<tr>
<th>id</th>
<th>skos:prefLabel</th>
<th>skos:altLabel</th>
<th>skos:related</th>
</tr>
</thead>
<tbody>
<tr>
<td>62</td>
<td>niod-terms Bezetting</td>
<td>Bezetting</td>
<td>..massavernietigingswapens ..militairewapens</td>
</tr>
<tr>
<td>2</td>
<td>gtaa 2156</td>
<td>bezetting</td>
<td>..onderdrukking,oorlogen ..bevrijding ..annexatie ..inzas</td>
</tr>
<tr>
<td>53</td>
<td>niod-terms Bibliotheeken</td>
<td>Bibliotheeken</td>
<td>..videotheken,archieven ..collecties ..manifestaties ..documenten boeken</td>
</tr>
<tr>
<td>54</td>
<td>unsure</td>
<td>bibliotheken</td>
<td>..massabijeenkomsten ..Partijdagen ..Manifestaties ..Ontmoetingen</td>
</tr>
<tr>
<td>55</td>
<td>niod-terms Binnenvaart</td>
<td>Binnenvaart</td>
<td>..religieuze diensten ..beloning ..masterclasses ..culturele manifestaties ..werkshops ..plechtigheden</td>
</tr>
<tr>
<td>2</td>
<td>gtaa 21609</td>
<td>binnenvaart</td>
<td>..trekvaarten ..schippers ..binnenvaartlopers ..waterweg ..binnenvaart</td>
</tr>
</tbody>
</table>

*CultuurLINK: result table showing the links between concepts found by an exact string match on the preferred labels.*

Because not all the concepts are mapped, we continue with these and try different string matching techniques and different attributes. For example, using the alternative labels, including stemming, and using fuzzy string matching (given a maximum edit distance). By considering labels of concepts that are not exactly the same we managed to find many additional links. Of course, this comes at a cost: more of these links contain errors. It depends on the use case how these errors are best detected. For example, hierarchical relations among the concepts can be used to filter out links for which the parents of the concepts are also linked. In this example the number of hierarchical relations contained in the vocabularies is too limited to effectively apply such a filter. Furthermore the number of links that are found with the additional blocks is in this case small enough to check by hand. The figure below shows the strategy after two additional string match blocks were added. Both blocks continue with the concepts for which no links were found in the previous block. This is done by connecting to the NOT output of the block.
**CultuurLINK: strategy with three string match blocks.**

At this stage more than half of the concepts from NIOD are linked to a concept from GTAA. By manually inspecting the concepts that are not linked it becomes clear that many of these do not occur in the GTAA as they are very specific topics for the NIOD Institute for War, Holocaust and Genocide Studies.

The links created between these vocabularies were used in the Linked Open Images application to recommend photographs from NIOD related to the videos from the Netherlands Institute for Sound and Vision.
6 Summary

Within COMSODE, Spinque extended its software stack to support advanced search application on top of ODN and Linked Data in general. The software is demonstrated with several applications. In the first year we developed a proof of concept to search the collection of the Rijksmuseum Amsterdam using background knowledge from DBpedia. We also created a search application for Czech restaurants that combined various data sources including information from the Czech inspection authorities. In the second year we developed a mobile application that provides integrated access to multiple cultural heritage collections.

The development of the applications have raised new requirements on the process of modeling search strategies in Spinque. In particular dealing with heterogeneity in Linked Data makes it difficult to assume that a strategy modeler knows the content and structure for all data sources. To better handle this situation we started the development of a data-driven version of the strategy editor. A first prototype is operational.

We started the migration of data preparation steps that were required at the indexing the phase. For example, data integration and the construction of text indices required pre-strategy configuration of a Spinque expert. Within COMSODE we have made the first developments towards real-time indexing that is automatically determined based on the strategy, reducing the need for a Spinque expert. We also developed a dedicated tool to support data integration. With LINK we applied Spinque’s Search by Strategy approach to the process of creating links between data sources. In the COMSODE pilot with the Netherlands Institute for Sound and Vision we successfully used LINK to align vocabularies in the cultural heritage domain. The resulting links are used in the Linked Open Images to support integrated access.

The current effort towards the end of COMSODE aims at the release of Spinque 2.0 platform and includes continuing the developments on the new strategy editor, real-time indexing and LINK.