



STERNA

Semantic Web-based Thematic European
Reference Network Application



METHODOLOGY FOR CONTENT ENRICHMENT

July 2010

Hans Nederbragt, Maarten Heerlien (Trezorix)

WWW.STERNA-NET.EU



STERNA

Semantic Web-based Thematic European
Reference Network Application



METHODOLOGY FOR CONTENT ENRICHMENT

July 2010

Hans Nederbragt, Maarten Heerlien (Trezorix)

WWW.STERNA-NET.EU

ABSTRACT

The STERNA project mainly focuses on enrichment of existing content of content holding organisations in the natural history domain. Therefore, developing a methodology on how to best integrate one's content into the STERNA information space is an essential part of the project.

This document is the outcome of that developing process. It describes in detail a six-step procedure on how to go about content selection, how to submit the necessary information to enable our technology providers to develop a domain-specific data model that incorporates all the required parts, and how to approach the actual content enrichment process.

In addition, this document contains three annexes that focus on data modelling in STERNA, on relevant examples of content items in the STERNA data model, and finally, on getting to know your way around the RNA Toolset, a set of versatile, easy-to-use tools that support various aspects of the content enrichment process.

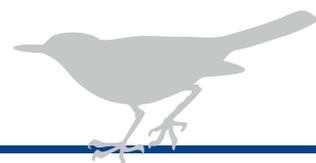


TABLE OF CONTENTS

Introduction	5
About this document	5
Relations to other activities and documents	6
Content enrichment activities (work package 3)	6
Technical issues (work package 4)	6
Evaluation of organisational and technical approach (work package 5)	6
Network extension (work package 6)	6
1 Step 1: description of use cases	7
1.1 What is the target group?	7
1.2 What is your audience looking for?	7
1.3 What kind of results do you want to offer them?	7
1.4 How does your use case relate to the bigger picture?	8
1.5 How do these use cases relate to the user scenarios?	8
1.6 Results of Step 1	8
2 Step 2: description of sources	9
2.1 What sources can serve the use case?	9
2.2 How can you access the data?	9
2.3 Do you own the source, or do you have permission to use it?	9
2.4 What is the quality of the data?	9
2.5 Results of Step 2	10
3 Step 3: evaluation of sources	11
3.1 What are the data models of your sources?	11
3.2 What are the appropriate available metadata?	11
3.3 What are the relations with existing reference structures?	11
3.4 What is the best modelling setup with regard to the findability system?	11
3.5 Results of Step 3	11
4 Step 4: evaluation of import methods	12
4.1 Connectors and conversions	12
4.1.1 Live connection (non-generic)	12
4.1.2 Scheduled conversion (non-generic)	12
4.1.3 One-time conversion (non-generic)	12
4.1.4 Connectors based on a standard protocol (generic)	12
4.1.5 Excel Generic connector	12
4.1.6 Adding new content and/or metadata by hand (RNA toolset)	13
4.2 Results of Step 4	13
5 Step 5: modelling, matching and mapping	13
5.1 Creation of and consensus on the general data model	13
5.2 Consensus on some central reference structures	13
5.4 Results of Step 5	13
6 Step 6: realization of enrichment	14
6.1 Describing the selected procedure in detail	14
6.2 Enrichment through connector or conversion	14
6.3 Results of Step 6	14
7 Evaluation of enrichment	14

TABLE OF CONTENTS

Annex 1: data modelling in STERNA	16
The RNA-architecture	16
Getting metadata into the reference layer	17
Reference structures	19
Relations through properties	20
Annex 2: examples of content items	22
Annex 3: knowing your way around the RNA Toolset	33
A: anatomy of the RNA Toolset	34
B: most-used operations in the RNA Toolset	37
Creating a new structure	37
Opening a structure	38
Creating a new item	38
Editing a content item	39
Adding an annex item to a content item	43
Adjusting view settings.	44
Glossary	46
Footnotes	47



INTRODUCTION

As a Best Practice Network project, STERNA is pioneering the semantic enhancement and integration of digital resources from different partners' databases based on Semantic Web standards and techniques. STERNA is a showcase project of using such semantic enhancement methods and the capability they provide to link, search and access content from distributed and heterogeneous databases in novel ways.

This process of semantic enhancement and integration of digital collections from the partners in the STERNA consortium is commonly known as *content enrichment*, an umbrella term for various methods to integrate distributed content into the findability layer. The findability layer is the RDF and SKOS-based metadata layer that serves as intermediate between this distributed content and the search interfaces that allow end users to search and retrieve this content as if it is stored in one single repository. It also allows to semantically enrich this content further by linking it to central reference structures such as taxonomies and thesauri, and also directly to content of other partners.

The STERNA Consortium aims to contribute to the objectives and realisation of the European Digital Library (EDL), to make Europe's cultural and scientific heritage accessible to all. To this aim, STERNA's primary contribution is to develop an infrastructure that enables appointed aggregators to collect content from communities that normally would not be able to deliver their rich resources to initiatives like Europeana due to lack of financial means or technical know-how. Therefore, one of the main focus points in the STERNA project is to develop a methodology for content enrichment that is cost-effective and foolproof.

A first version of this methodology was developed in fall 2008 and has been introduced to the STERNA partners prior to the workshop on content enrichment held on 4-5 December 2008 in Leiden, The Netherlands. This first version of the methodology was serving mainly as a guideline document for the preparatory work that needed to be done with regards to content evaluation and selection. After the workshop, the methodology was further developed and validated by the consortium members during the first cycle of content enrichment, where small data sets were enriched. After this first round, the methodology was further refined and now serves as primary input for the second cycle of content enrichment where partners will enrich their full collections of content. This second cycle of content enrichment started in mid-April 2009, and will continue until summer 2010.

About this document

This document serves as basic guidelines for completing the process of content enrichment in a cost-effective way to both current and future partners of the STERNA network. To this end, a six-step methodology is described in detail that encompasses the entire content enrichment process, from envisioning the possible user groups for each collection to the actual enrichment of the content itself. At the end of each step in the methodology, a list of expected results is added that should have been accomplished after completing this step.

In addition to the six steps of the content enrichment (CE) methodology, there are three annexes to this document, each containing specific information that should aid you, as a content providing organisation, in making the right choices at one point or another during the process of content enrichment.

The first annex contains the STERNA data model in its current form. The purpose of each item type is explained in this annex, giving content partners an insight into the basic building blocks of the data model.

The second annex contains a collection of examples of content items, giving you a rough idea of what the various types of content items may contain with regard to the predicates and their possible values within those content items.

The third annex is a basic manual to the RNA Toolset, a set of easy-to-use tools that provides you with the means to create and edit metadata, reference structures, rich text articles et cetera, and to semantically integrate your own content with the content of the other partners. This annex will help you finding your way around the tool set as it focuses on the basic features, followed by a description of several often-used operations with respect to content enrichment through the RNA Toolset.

Relations to other activities and documents

As content enrichment is one of the central aspects in STERNA, this methodology is closely related to several other activities and deliverables in the project:

Content enrichment activities (work package 3)

STERNA is mainly a content enrichment project, and the actual task of enriching content is carried out in work package 3. The objectives of this work package are, amongst other things, to enrich selected content with metadata to make it fit for distributed querying and to attach and edit reference structures. This methodology for content enrichment functions as a guideline to help partners but also new members to accomplish this task properly and effectively.

Technical issues (work package 4)

While working through this methodology, you are expected to make several choices that will affect the outcome of the content enrichment process regarding the specific method of content enrichment that suits you or your separate collections best. Some of these choices concern technical issues. A good understanding of the STERNA distributed Reference Network Architecture will aid you in making these choices. The annexes in this document, especially annex 1 will help you to gain this understanding.

Evaluation of organisational and technical approach (work package 5)

As explained shortly in chapter 7 of this document, the methodology for content enrichment will also be subject to a thorough evaluation process. The methodology will be evaluated from a technological, organisational, economic and user perspective. The strategy for evaluating the STERNA approach in technical or organisational terms has been described in more detail in the internal deliverable, DEL 5.3.1 – Evaluation methodology for the STERNA approach, which may be obtained on request.

Network extension (work package 6)

Extending the STERNA network to at least twelve new participants by the end of the third year is one of the performance indicators of the project. This is the subject of work package 6. The methodology for content enrichment supports this goal by developing a method which not only suits the current partners but which will also provide future members of the STERNA network with a road map to successful content integration and semantic enrichment.



1 STEP 1: DESCRIPTION OF USE CASES

Joining STERNA as a content provider means that you will semantically integrate your content into an innovative digital library, built with state-of-the-art semantic web technology. As birds are the central theme with respect to the content that will be made accessible through STERNA, we assume that you have one or more collections of content related to birds. From the STERNA point of view, what do we consider a collection? Below are a couple of examples:

- A set of articles with illustrations (for instance texts, photographs, illustrations and maps taken from a book on seabirds)
- A set of photographs or illustrations of birds
- Web pages, for instance a selection on birds, taken from a natural history website
- Collection registration records, for instance object descriptions of the bird collection of a museum
- Audio files with bird sounds
- Field recordings
- A digitized collection of stamps with images of birds
- A poem collection with birds as a theme
- Et cetera

To make it easier to decide how to make these collections ready for use within the STERNA project, you should start by describing a *use case*, i.e. an outline of how you want your content to be found and by whom.¹ This will help you to decide what kind of metadata is necessary to make your content findable and accessible through STERNA. These decisions in turn will advance the modelling process, which is described in the following steps of the procedure.

Your use case description helps the “modelling technicians” to understand what you want with your data. Without such descriptions they might make decisions that are correct from a technical point of view, but which may have an effect on the findability and presentation of your content that is the exact opposite of what you had in mind.

To help you with writing your use case(s), you should use the following four questions as guidelines.

1.1 What is the target group?

Define the audience or audiences that you primarily want to reach. Possible answers to this question could range from “anybody” to “birdwatchers” or “primary school children”.

Deciding on specific target groups can have an immediate effect on modelling. If, for instance, it is important that different collections can be related to specific end users, it is necessary to have some kind of metadata added to all content, such as the Dublin Core property *dc:audience*.²



¹For footnotes, see page 47.

The answer to this question is – in a more indirect sense – also very important to be able to answer the next question: what are your target users looking for?

1.2 What is your audience looking for?

Suppose you have a collection of image files of paintings of birds. Let's say that two kinds of user groups may be interested in these pictures: people who are primarily interested in biological aspects and people who are primarily interested in aspects related to art. If you want to serve the first user group, then properties like “scientific name” and “habitat” may be important. But if you want to focus on the second group you would have to apply properties like “title”, “artist”, “art period” and “painting technique”. And if you want to make your collection findable for everybody, then you have to add as much metadata as you have or can get.

To give you an idea of what kind of content users in a specific user group may look for, we have developed four *user scenarios*. These user scenarios, each focusing on a different possible user group, try to describe what kinds of content a user from each of these user groups may want to find and which search paths he or she would most likely want to use to find that content. As a reference, the four initial user scenarios developed for STERNA are available on the project website (<http://www.sterna-net.eu> >> Use cases).

1.3 What kind of results do you want to offer them?

Apart from metadata that you want to supply because you think they fit with the user group(s) you have selected, there also may be metadata that you need because you want to present the results of queries from the end user in a certain way. The most important guideline here is:

What you want to get out of STERNA, first must have been put into it.

Suppose, for instance, that you want to submit a collection of bird photographs to STERNA. If later in the project you would ask your own web developer to design an interface on your own website to search the STERNA collection where an end user specifically can search on the creator of the photographs, this will only work if there is metadata for your photographs containing a specification like the Dublin Core property *dc:creator*, with reference to the photographers. In other words: a certain degree of planning ahead is advisable when deciding on what kinds of results you want to offer to users of STERNA.

1.4 How does your use case relate to the bigger picture?

In light of the previous example, it is even better to add the *dc:creator* metadata field to all photographs in the STERNA project, as well as to the paintings from the previous example.

Now suppose the STERNA partners decide that in their to-be-developed user interface they want the end user to be able to differentiate between queries for artwork of painters and that of photographers. Then again, it is necessary to plan ahead in modelling and make two specifications of *dc:creator*, which could look as something like *dc:creator/rna:painter* and *dc:creator/rna:photographer* (the “rna:” indicates that it is something like an RNA-type of specification).

Talking about planning ahead brings us almost inevitably to something like “the big picture”. This is a difficult but exciting issue, because one can argue that on the web there is no “big picture” anymore, or that there are so many big pictures that they completely lost their “big-picture-quality”: “everything is miscellaneous”.³ Should all STERNA partners freely submit the collections they want to submit, or should there be some filtering, for instance by using the user scenarios as guidelines? The first option is maybe truly “thinking Web Three”, giving room for the unexpected and so on, but the second option might be more practical from a project management point of view. We propose an in-between solution: try to categorise material you want to submit in relation to the four user scenarios. What fits passes the test immediately and can be described into use case descriptions. Collections that do not fit might eventually be useful for one or more extra user scenarios, and finally, all that is left over will be discarded.

1.5 How do these use cases relate to the user scenarios?

The use cases as described above have a much narrower scope than a user scenario.

The use cases will be used by the more technically oriented modellers, to relate what they are doing to the original meaning and purpose of the creator of the database. Having such a recipe makes their work easier. If the use cases are clear enough, they do not have to ask all the time: “what do you mean by this?”

The purpose of the user scenarios is much broader. They deal with the meaning of the whole STERNA project, and what users might get out of it.

1.6 Results of Step 1

After completing Step 1 of the methodology for content enrichment, your results should be:

- The name and a short description of each collection that you want to add to STERNA.
- A short description of the most important target group(s) that you assume will be interested in your content, for each collection that you want to submit.
- A short description for each target group of what you think they are looking for in the specific collection.
- A short description for each target group of what you want to present to them as results when they are searching your collection.
- A short description for each collection of what you think is the relation with your own use case or the four originally developed user scenarios.

To support you in these tasks, we have developed an MS Excel template that you can fill in step by step. The template can be downloaded from the project website as well (<http://www.sterna-net.eu/> >> Content Enrichment >> Template for content enrichment). The results for step 1 in the methodology should be described on the *step 1* tab page of the template.



2 STEP 2: DESCRIPTION OF SOURCES

In the previous step we used a collection as a starting point to describe one or more related use cases. In this section we take these use cases as a starting point to define more precisely what part of the collection should be used and what other resources might be useful for the use case.

2.1 What sources can serve the use case?

As far as content is concerned, there are many types of sources that you could use for your use case. We mentioned some in the previous section: sets of articles, photographs, web pages, object descriptions, audio files, field recordings, et cetera. But at this point you should also look at the more technical aspects of your selected sources. For instance, what formats are used?

- Your sources can consist of all sorts of files, which can have many formats.⁴
- Your sources can also exist as web pages, which might be available either as separate files (formats: html, xhtml, xml) or via a website.
- And, of course, a source can also be a database, or even a spreadsheet table.

2.2 How can you access the data?

If your source is a collection of separate files, then it is important to investigate where exactly these files are stored and how they can be accessed:

- What is the address of the location where they are stored?
- Do you need a password to get to this location?
- Are the files themselves protected in any way (for instance by login name and password)?
- How can the files you want to use be selected (for instance: can you select them on file names, file dates, metadata, a full text query, et cetera)?
- What are the formats of the files you want to use (for instance: Word, PDF, jpeg, MS Access database)?

Your source might also be part of a website. In that case it is important to know how the web pages can be accessed and what they look like:

- Are the content items available as downloadable files, or can they only be accessed via the website?
- If they have to be accessed via the website, can they be accessed via the http-protocol, or via a web service?
- Which file formats are being used?

And finally your source might be a database. In that case it is important to get to know through what database management tools the data can be accessed:

- Can the data be accessed by way of a regular database management system (for instance MySQL, MS Access, or Oracle)?
- Must the data be accessed by way of a specialised database management system (for instance collection registration systems like OpenCollection or Adlib) that “sits on top” of a regular database management system?
- Or can the data be accessed using specialised functionalities of a specialised database management system (for instance the OpenArchive functionality in Adlib)?

2.3 Do you own the source, or do you have permission to use it?

This is a very important question, as it has to do with intellectual property rights (IPR). This is not only a matter of “owning” the content itself, but also of owning the metadata accompanying the content, and of having permission to use embedded content (like photographs in an article) or linked content (like content “behind” a link in a webpage).

When you do not own a source, but you do have permission to use it (like content in a website) then it is important that you arrange both technical access issues (see paragraph above) and intellectual property issues at the same time.

If you are not certain about who owns the intellectual property rights of a specific source and about what you may and may not do with that source in terms of publishing and digital accessibility, please consult a legal consultant on this matter. Clearing IPR issues regarding a source is the sole responsibility of the partner who provides that source. Please note that violating intellectual property rights may result in serious financial claims from the party that holds the IPR.

2.4 What is the quality of the data?

In daily practice, the quality of the data in databases is the most frequent cause of disappointment for the owners of those databases. Often, they have a good idea of what the database is about and how the information in it can be of use, but when put to actual use there are often many details that make a less sunny picture. A couple of examples:

- The database does not have all the fields that you thought it had. This might seem strange, but this happens quite often. It may result in disappointing results when querying the database, as it might be impossible to get information out of the database in the way you had envisioned it.

⁴For footnotes, see page 47.

-
- The database contains incomplete records; record fields have been left empty. This is another common problem with databases, because when working on database records often “for the time being” only a couple of fields are filled in – the rest to be “filled in later”, but in practice they are forgotten forever. The result is that, although the right fields are available, the records with empty fields will never be found.
 - The data are there, but they do not meet the needs of the end users. For example: a large taxonomic database, with very interesting content, but where the animals and plants in it can only be accessed through their scientific name. The “general audience” the database is meant for does not know these scientific names, so there is no way users from this audience can get to the interesting content.
 - All the right fields are there, they are all filled, the data are of good use to the end user, but there is no consequent use of the data. For example: the name of one of Naturalis' experts on dragonflies is Jan van Tol, which can be written as: *Jan van Tol, J. van Tol, J. v. Tol, Jan v. Tol, Tol, Jan van*, et cetera. Another example: in a certain database *twelve* different ways were used to write dates. The effect of this is that only part of the dates will be found, unless you instruct the end user about all of the idiosyncrasies that are present in the database.

Important: these examples are by no means to argue that the Reference Network Architecture setup that is used in STERNA will do miracles with “bad” databases. On the contrary, they are to warn you that when you run into problems like the examples above, not acting on them will result in disappointment both in finding the data as well as in the way it is presented through STERNA. So, look at the quality of your data before you create too big expectations and either take countermeasures to improve the records in your database, or adjust your use case.

2.5 Results of Step 2

After completing Step 2 of the methodology for content enrichment, your results for each collection should be:

- An overview of the formats used for the collection (e.g. Word files, HTML web pages, database).
- A description of how and where the collection can be accessed (for instance: as a database on DVD, via a web service at www.example.org, via ftp-server at www.ftpaddress.org with username “user” and password “pass123”).
- A description of any IPR issue that might be of influence on the use of the data in the STERNA project.
- A description of the measures that have to be taken before the database is usable for STERNA, with the type of actions you want to take and on which data (e.g. *cleaning up date field using the ISO 8601 yyyy-mm-dd format*). Add expected date of completion.

The results for this chapter should be described on the step 2 tab page of the template (see 1.6).



3 STEP 3: EVALUATION OF SOURCES

In step 3 of this methodology for content enrichment, you will describe the more technical aspects of the collection(s) you have selected. At the end, a first concept is made of the modelling of your collection for the findability system of STERNA. This last part is done in cooperation with our technology partners, Trezorix (www.trezorix.nl) and NCB Naturalis (www.naturalis.nl).

3.1 What are the data models of your sources?

The most important questions are:

- What type of entities (item types) are important in your source?
- What are important properties of these item types with respect to how you want to make them findable?
- How do these properties relate your entities to other entities?

These three questions have to do with the basic building blocks of modelling for the semantic web: **item types**, **predicates** (as the typification of properties), and **domain and range** (to determine linking possibilities between different item types). This is explained in Annex 1: data modelling in STERNA.

Suppose you have the entities “paintings”, “painters”, and “place of exhibition”, then there are relations between paintings and painter and also between painting and place of exhibition. The three questions above can be answered by looking how the original data is modelled (use spreadsheet exports for quick and easy overviews) and translate this to the data modelling in the Sterna environment.

3.2 What are the appropriate available metadata?

Something to take in account here is the importance of the different fields in the data model relative to the kind of findability that follows from your use case description. If, for instance, your collection exists of paintings with birds as subject, and your target group is birdwatchers, then the field “painting technique” is probably of little importance.

3.3 What are the relations with existing reference structures?

Another important aspect is the relations that certain fields may have with existing reference structures. For example, a field “scientific name” might very well be related to a specific taxonomic thesaurus. Or the

field “country” with the ISO 3166 standard for country names.

These structures do not necessarily have to be (inter-) national standards or commonly accepted thesauri. Also “homemade” vocabularies are important. These may vary from a text list that is used as a reference when fields are filled in by hand, to selection lists that are used as a more controlled way of input for fields. If you have a collection that has a relation with a reference structure of your own, you should describe this structure as well and send a sample of the database.

3.4 What is the best modelling setup with regard to the findability system?

When all the above data are collected a first concept of modelling can be made. The central question in this task is how your collection can be related to the overall findability structure. This modelling will be an iterative process between you as the domain expert and Trezorix and NCB Naturalis, in which we will analyze your data and make a proposal, after which you review and correct the proposal. Examples of the current data model (version 2) are provided in Annex 2: examples of content items.

3.5 Results of Step 3

After completing Step 3 of the methodology for content enrichment, your results for each collection should be:

- A list of item types that are appropriate to represent the data you want to get to the reference layer.
- A list of predicates (typifications of properties, like “hasAuthor”) that are appropriate to make the links between the data you want to get to the reference layer.
- A representative sample of the database, in the form of a (some) spreadsheet(s).
- A description of the reference structures – homemade or commonly used – that are used for data entry of fields in your database.
- A representative sample of, or entire homemade reference structures.

The results for this chapter should be described on the step 3 tab page of the template (see 1.6). Send the template with your results for steps 1 to 3 to Trezorix (sterna@trezorix.nl) or NCB Naturalis (sterna@naturalis.nl) together with the database samples and samples of any homemade reference structure you might be using.

4 STEP 4: EVALUATION OF IMPORT METHODS

In this step, the method that will be used to put your data in the findability layer of the STERNA information space will be decided upon. There are several methods to do this and it depends on the nature and the size of your dataset(s) which method is most suitable for each of your collections. The collections will be evaluated by Trezorix with respect to the type of enrichment that will fit best to the specific collection on the one hand, and to the overall requirements of the STERNA project on the other hand. This is done together with the collection owner.

4.1 Connectors and conversions

One way of enriching your content is through the use of a data connector or a conversion. The term data connector is an umbrella term for various functionalities that facilitate the connection between the database layer and the findability layer. Below is an overview of the various connector types and conversion methods and the types of datasets they can best be applied to.

4.1.1 Live connection (non-generic)

In this method, data is being collected ad hoc from a legacy database when a query is sent to it. The database needs to be equipped with a layer that translates the RDF format to the format that is used in the database and vice versa. The live connection can best be applied in situations where:

- The database has a very dynamic nature: changes in the data occur often, for instance because the source database is still under development.
- Only the original database may serve as a validated source.

4.1.2 Scheduled conversion (non-generic)

This form of conversion is being used in situations where data in legacy databases needs to be kept up-to-date and needs to serve as the original source. Conversions can be executed automatically according to a time schedule or they can be triggered by mutations in the legacy database. The scheduled conversion can best be applied in situations where:

- The database has a somewhat dynamic nature: changes in the data occur regularly,
- Only the original database may serve as a validated source.

4.1.3 One-time conversion (non-generic)

This form of conversion is being used when the data in a legacy database, the original database, does not have to be edited anymore or if the converted data is being edited further in the RNA environment. The one-time conversion can best be applied in situations where:

- The original database is not being developed any further.
- The owner of the database wishes to continue to work on the RDF representation of the database.
- The database requires a lot of one-time corrections by hand, which can be done in the conversion process.
- The database includes many interpretations that are hard to translate to reliable queries (an example of this is inference of hierarchy).

4.1.4 Connectors based on a standard protocol (generic)

If a connector is created for a database that needs to be approached through a standardised protocol, this connector can be used for every database that can be queried through this particular protocol. Examples of such protocols are the OAI protocol⁵ that is often being used for accessing archival databases, and the TAPIR protocol⁶, which is often applied in the natural history domain. The connectors based on a standard protocol can be applied in situations where the database management system makes use of a standardized protocol for exchanging data.

4.1.5 Excel Generic connector

The Excel Generic connector has been developed for importing table data into an RNA environment. The Excel Generic connector works by creating an MS Excel template that contains data from a legacy database. Outside of the toolset, the data is entered in the a spreadsheet template. Because links can be imported as well through this connector, parts of a relational database can also be imported into the RNA environment this way. The Excel Generic connector is the most common tool to import data from a database into an RNA environment.

4.1.6 Adding new content and/or metadata by hand (RNA toolset)

Another possibility is to create metadata and content directly in the STERNA information space, without converting data from legacy data collections. Some considerations that may help you to decide whether *metadata should be added by hand*:

- The collection does not have metadata (for instance: the collection exists of bitmaps of pictures, with no accompanying metadata yet).
- The whole collection has to be built from scratch (for instance a list of literature references, which consists at the moment of start of some Word documents).

4.2 Results of Step 4

After this step, a strategy on how the data from your legacy datasets will be made part of the findability layer in the STERNA environment and if and how further enrichment of your data will take place once it is part of that environment should be outlined.

5 STEP 5: MODELLING, MATCHING AND MAPPING

As mentioned in paragraph 3.5, when a modelling concept for every separate collection has been made, the modelling with regard to the whole set of collections will commence. In particular, this means that a more or less general data model will be configured as a starting point.

5.1 Creation of and consensus on the general data model

To ensure optimal integration of the content collections of all of the content providing partners and to allow the end user to search through these collections as if they are one big collection, a data model is created where the findability and presentation data of all of these collections can be fitted in.

Your input at this point is crucial as you are the best candidate for judging whether or not the data in your collections can be fitted into the model and whether or not it will serve the user group(s) you have envisioned. In reviewing the proposed data model there are some important questions:

- Are there certain item types missing?
- Do the properties (predicates) allow for the planned relations between items?
- Are the relations between the various item types complete and logical (check domains and ranges)?⁷

5.2 Consensus on some central reference structures

In paragraph 3.3 we described the importance of the relation between your data and common or local reference structures that play a role as controlled vocabularies. A certain level of consensus amongst the partners is needed on the reference structures that are used this way. Some aspects are important here:

- To what extent do the data match with the concepts that are present in the reference structure?
- Are important concepts missing?
- Is there an alternative?

5.4 Results of Step 5

At the end of step 5 there should be a data model that the partners agree upon as well as a set of reference structures that are commonly accepted by the partners.



⁷For footnotes, see page 47.

6 STEP 6: REALIZATION OF ENRICHMENT

In this chapter we will focus on how to proceed with the enrichment process and we will introduce you to several tools to enrich your data, which are explained in further detail in annex 3 of this document.

6.1 Describing the selected procedure in detail

Enrichment is simply adding extra metadata to your content or – which is more common – cleaning up existing metadata.

Below are some general suggestions for drafting an enrichment work plan, which is a useful first step:

- Make a map to determine which fields from the legacy dataset should go where in the STERNA data model (see also paragraph 5.3).
- Make an analysis of which data have to be added or cleaned up.
- Decide whether this can be done best before or after importing your data in the RNA environment, and how.
- Also decide how changed metadata can be (re-)linked to other content items.

An example of incomplete data are empty fields in a table column. Very common examples of data to be cleaned up are the use of different data formats in a column, like dates (01-02-1960, February 1st 1960, etc.) or names (Charles Darwin, Darwin, C.).

6.2 Enrichment through connector or conversion

When you enrich your data by connector, in whichever form (see step 4), the enrichment will be done in collaboration with the information technicians of Trezorix, as they build the connectors and, in most cases, perform the upload of the data to the STERNA information space. Specific technical details in these cases should be discussed with Trezorix as well as the specific methods to match data in the legacy datasets with concepts in the central reference structures.

6.3 Results of Step 6

The results of this final step in the methodology for content enrichment should be:

- A collection of content that is findable and presentable through the STERNA environment, in the way that was envisioned in the use case that was drawn up for the collection in step 1, and that is semantically linked to content of other partners either directly or through concepts in central reference structures.

7 EVALUATION OF ENRICHMENT

After the completion of the content enrichment procedure, the process will be evaluated to determine if the results of the process meet up with the objectives that were envisioned and described in the use cases at the beginning of the whole process. The central question of this evaluation with respect to the content enrichment itself will be whether or not the end user will be able to find in STERNA what he/she is looking for and in the ways that fit best to his/her specific needs and wishes.

Another object of evaluation will be the central reference structures that are used in STERNA and if and in what way they serve their purpose as the semantic glue between the content of the various partners. With respect to the reference structures, the relation between concepts in the structures and the number of content items they are connected to will, amongst other things, be evaluated.

A third focus point of the evaluation will be the tools that are used for the process of content enrichment. Central issues here will be whether or not these tools meet the requirements of the content providing partners to be able to comfortably steer through the process of enriching their content.

The specific outline however, of these evaluation questions and the methods that will be used to answer them will be addressed in another STERNA deliverable, to be produced as a result of the work done in work package 5 of the project.⁸



ANNEX



ORIOLES BRONZED, *Oriolus*

ANNEX 1: DATA MODELLING IN STERNA

The RNA-architecture

This document explains the modelling possibilities of the new version of the RNA Toolset, version 4. RNA is an acronym for *Reference Network Architecture* that was coined in the RNA project⁹, the predecessor project of STERNA which focused on developing best practices for working with dynamic knowledge systems.

A dynamic knowledge system is defined as an information environment that enables users to comfortably add new content, which is integrated instantly and as automatically as possible into a findability system based on semantic web technologies (providing more than just *full text search*). This flexibility is achieved by linking content in a reference network through the use of metadata.

The RNA architecture is a web-based architecture that makes it easy to connect knowledge sources on the web and to make the heterogeneous content in these resources retrievable in a comfortable and unambiguous way.

In the RNA architecture, three layers are distinguished:

The bottom layer is a *data layer*. This layer contains data from databases, file systems and web pages. These data can be accessed through the Internet. The bottom layer represents the *supply* of information.

The top layer is an *application layer*, containing applications that want to make use of the data in the bottom layer. In this application layer the *demand* for information is determined.

These two layers are connected through an intermediate third layer, existing of a *network of references*. This layer contains linked metadata from the data layer and hierarchically arranged structures in which these metadata are stored. The reference layer serves as a mediator between the supply and demand sides, by combining high-quality retrieval functionality with a detailed overview of the available data.

The reference layer, together with the interfaces that are available for communication and exchange with both other layers, is called an **RNA environment**. The tools that were developed to manage and extend the reference layer in an RNA environment are called the **RNA Toolset**.¹⁰

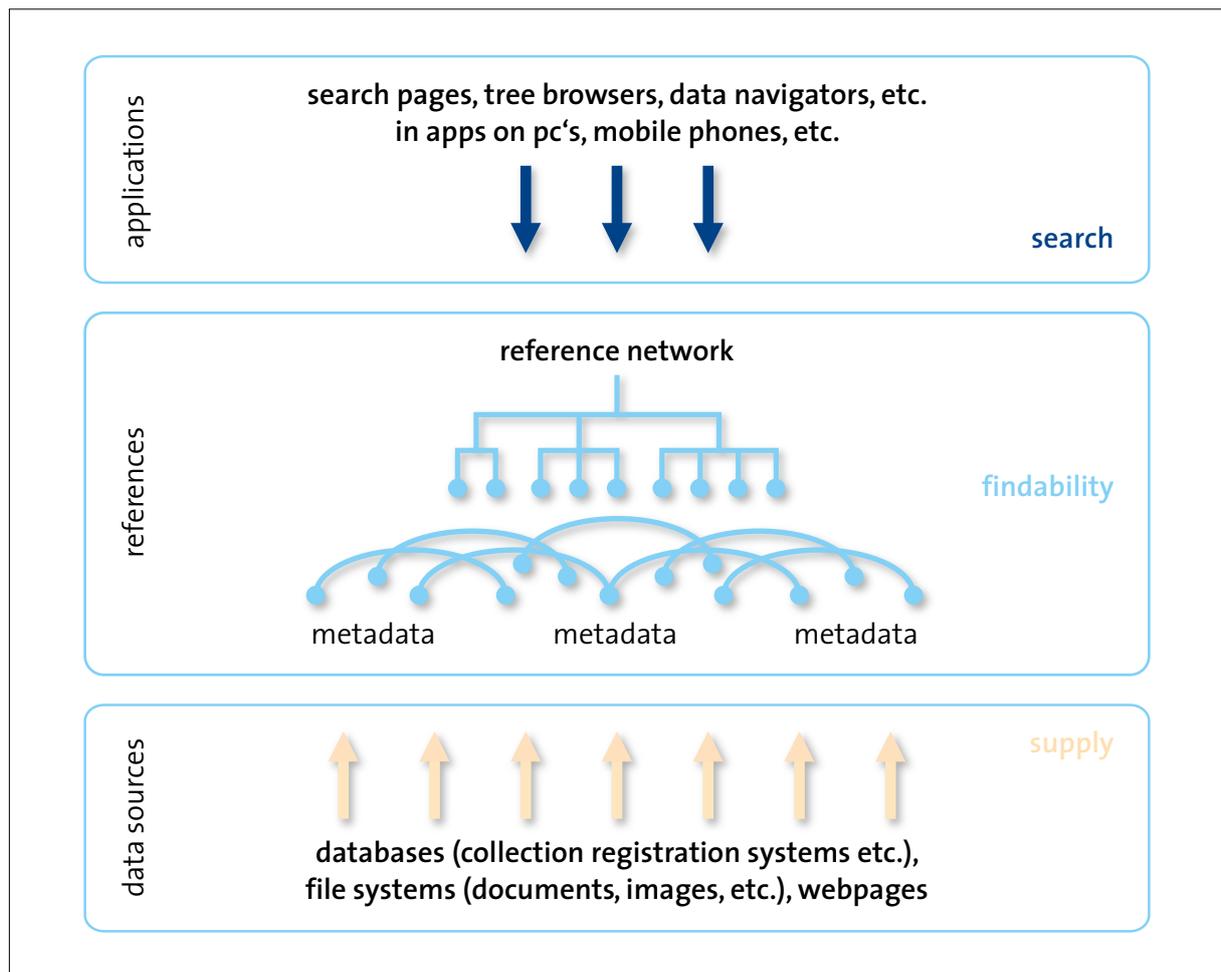


Figure 1: The three layers of the RNA architecture.

Getting metadata into the reference layer

We will now give an example of how metadata are generated from a source in the data layer.

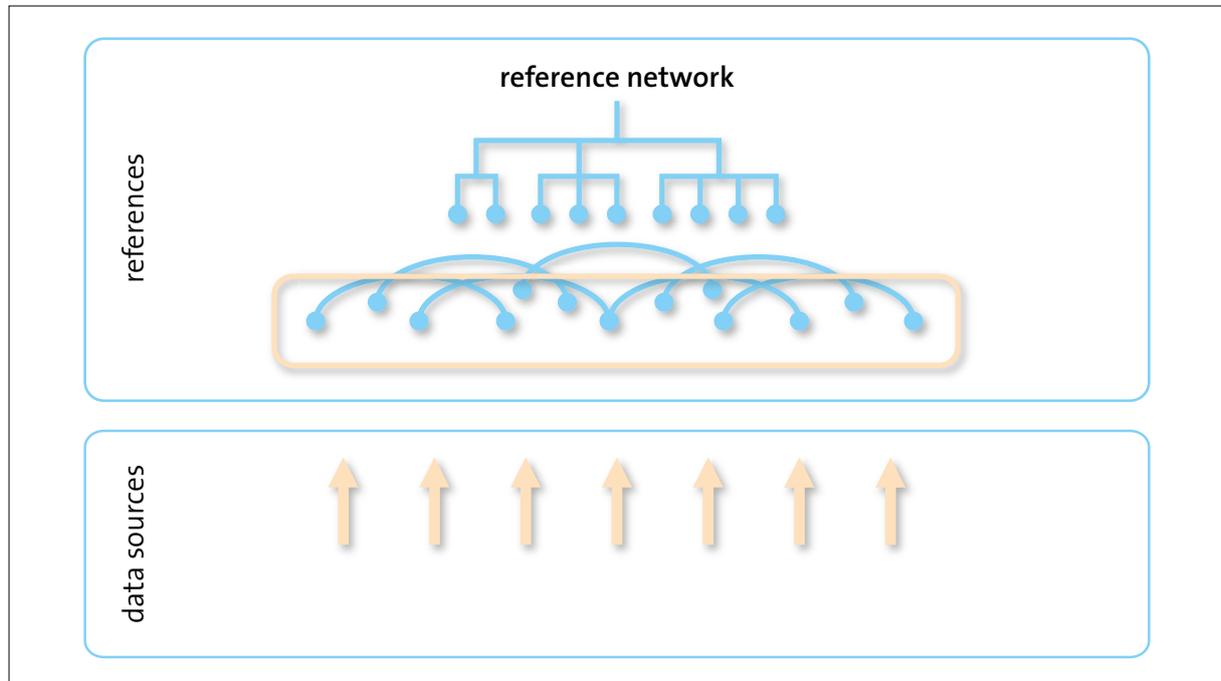


Figure 2: Generating metadata from a data layer source.

Below is an example of a record that could be found in a database that is part of the data layer. The record describes some properties of the book “On the Origin of Species”, written by Charles Darwin.

title:	On the Origin of Species
author:	Charles Darwin (1809-1882)
creation date:	1854 - 1859
publisher:	John Murray
publication place:	London, United Kingdom
publication date:	November 24, 1859
language:	English
audience:	biologists, general audience
pages:	502

Figure 3: Metadata of a book

The record describes an entity: the book. But it contains references to other entities as well:

- to the author (a person)
- to the publisher (a company)
- to a city and a country as place of publication (both locations)
- to the language in which the book is written
- to two roles that indicate the kinds of readers the book is meant for

title:	On the Origin of Species	
author:	Charles Darwin (1809-1882)	actor/person
creation date:	1854 - 1859	
publisher:	John Murray	actor/company
publication place:	London, United Kingdom	town, country
publication date:	November 24, 1859	
language:	English	language
audience:	biologists, general audience	roles
pages:	502	

Figure 4: Description of a book entity.

The record also makes reference to two entities that contain some of the entities we just mentioned:

- a *creation event*, related to the creation of the book
- a *publication event*, containing properties regarding its publication

creation event:	title:	On the Origin of Species	
publication event:	author:	Charles Darwin (1809-1882)	actor/person
	creation date:	1854 - 1859	
	publisher:	John Murray	actor/company
	publication place:	London, United Kingdom	town, country
	publication date:	November 24, 1859	
	language:	English	language
	audience:	biologists, general audience	roles
	pages:	502	

Figure 5: References to other entities.

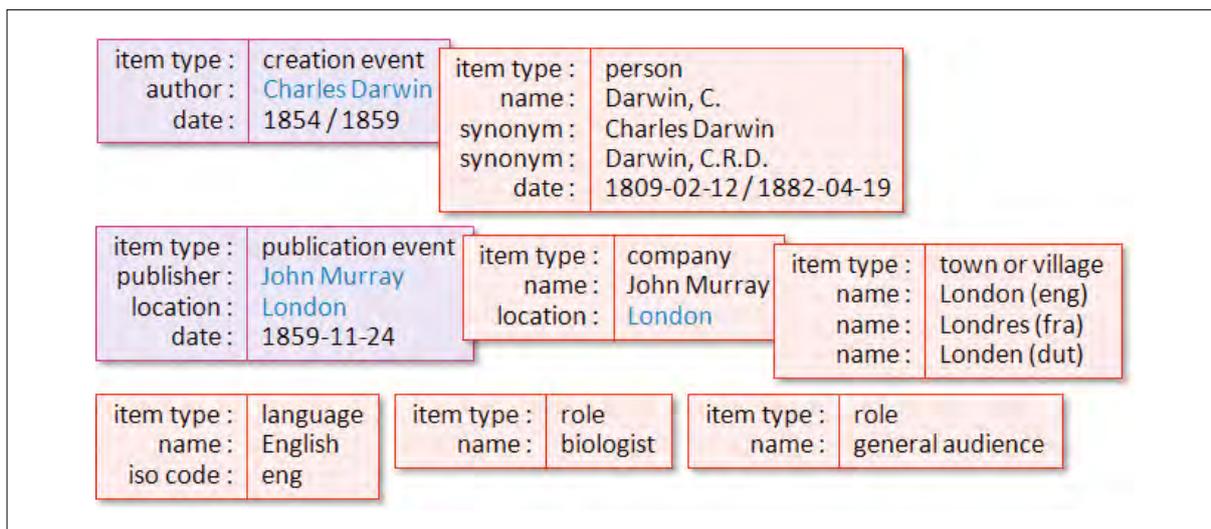


Figure 6: Metadata regrouped into new entities.

All these data, that may be important for findability, will be moved to the reference layer as metadata, and they will be regrouped into new entities, as depicted in figure 6.

The blue-coloured entries in the right part of these items are links to other entities, such as the *person* item with data on Charles Darwin, the *company* item regarding John Murray, and the *location (town or village)* item for London.

The black-coloured entries in the right part of the items are *literals*: for instance a text (*string*), number or date. These entries do not refer to other items.

In the *person* item for Darwin a *preferred name* is used: Darwin, C. This name is suited for a common way of searching for content, based on surnames. In addition there are two synonyms. Each of these names, when used as a search term, will lead to the item about Darwin. In the same way it does not matter whether you search for London using the English, French or Dutch spelling.

Note that some of the extra data in the example are not gathered from the original record, but were apparently retrieved from somewhere else.

When a more accurate level of findability should be realized, extra entities can be added. In the example below, an extra *birth event* is added.

This item again has references to other items, such as Darwin's place of birth, his father and his mother. In principle this can go on until the desired findability is achieved.

The items that are created in this manner in the reference layer are called *content items*.

Instead of extracting them as metadata from the data layer, these content items can also be created directly in the reference layer. In several projects, large numbers of webpage-like content items are created this way on an editorial level. In addition to the record-like data as in the examples above, these content items also contain web content such as texts and images. In this type of workflow, the RNA Toolset serves as a content management system (CMS). The content items that are produced this way are both part of the data layer (the web content) and the reference layer (the metadata of the web pages).

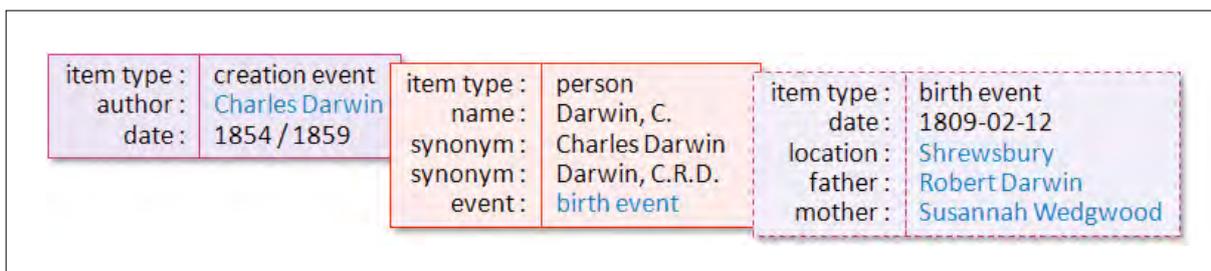
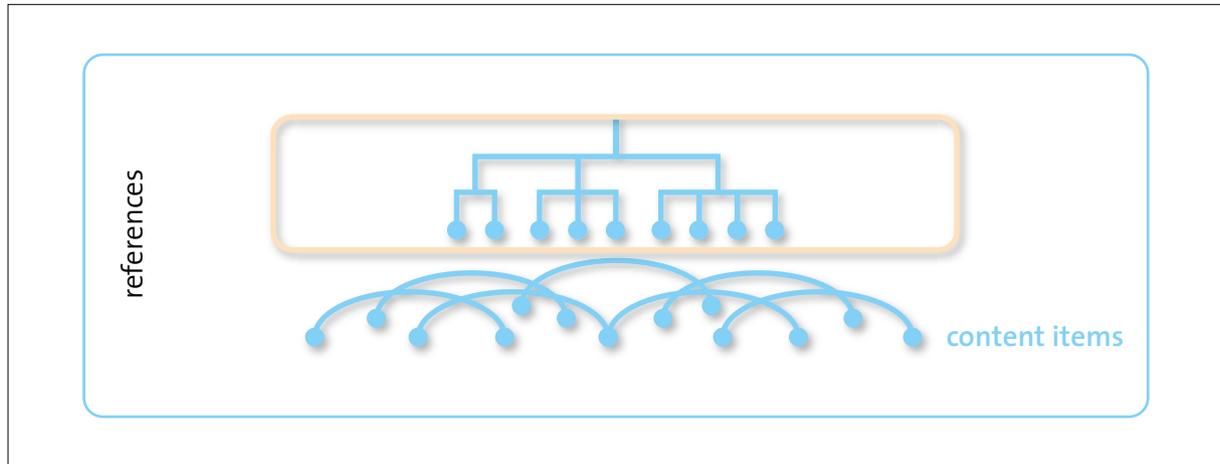


Figure 7: Adding extra entities (for example, birth event).

Reference structures

The *content items* that are being created in the reference layer in the previously described way are placed in hierarchically ordered structures, as exemplified in the figure below.

Figure 8:
Placing content items into a hierarchical architecture.



This arrangement in structures can be done in accordance with several different criteria. We will clarify this with the help of the previously used example: the record describing the book *On the Origin of Species* by Charles Darwin.

title:	On the Origin of Species	
author:	Charles Darwin (1809-1882)	actor/person
creation date:	1854 - 1859	
publisher:	John Murray	actor/company
publication place:	London, United Kingdom	town, country
publication date:	November 24, 1859	
language:	English	language
audience:	biologists, general audience	roles
pages:	502	

Figure 9: Arranging entities into hierarchies: the Darwin example.

In this example there are four aspects that may play a part in arranging items into reference structures:

- File management: this regards convenient management of content items, and also authorization in an RNA environment is handled through structures.
- Hierarchical inference: the reference to London obviously concerns the city of London in the United Kingdom, and not a London someplace else.
- Concept completion: searching for synonyms and language variants will lead directly to the concept they are part of.
- Inheritance: the children of the concept *actor* will inherit properties from their parent item, and can add their own properties to this inherited set.

The figure below provides examples for arranging entities into structures. In addition to these criteria, reference structures can also be set up according to the arrangement of commonly used thesauri.



Figure 10: Examples of arranging entities into hierarchical structures.

Relations through properties

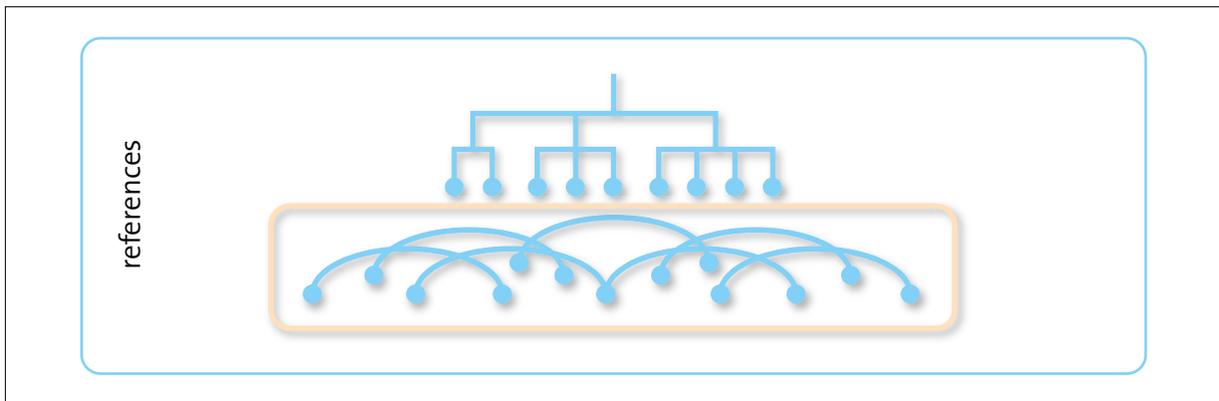


Figure 11: Creating relations between content items.

In the previous chapters we have demonstrated how metadata can be generated from resources within the data layer, how these metadata are grouped into content items in the reference layer and how these content items can be placed in reference structures.

Now we will demonstrate how *direct relations* between content items can be created.

These direct relations are primarily created to represent certain properties of content items. This is done with the use of a *statement*. A statement consists of three parts:

- A **subject**, i.e. the item something is stated about. In the example below this is the content item “On the Origin of Species”.
- A **predicate**, i.e. the typification of the property. In the previous example, this is the predicate “hasAuthor”.
- The **value** of the property. In the example this is Darwin. This may either be a link to a content item with “Darwin, C.” as its name, or a *literal*, the string of characters that form the name.

As shown in figure 12, the left content item is of the type *book*, since it is connected to the **item type** of the same name. The content item on the right is of the type *person*.

The collection of item types that can make use of a specific predicate is called the **domain** of that predicate. So, because the item type *book* is in the domain of the predicate *hasAuthor*, a content item of the type *book* can use this predicate.

The collection of item types a predicate can refer to is called the **range** of that predicate. So, because the item type *person* occurs in the range of the predicate *hasAuthor*, a content item of the type *book* can be linked to a content item of the type *person* through this predicate.

This construction, frequently used in semantic web applications, is an essential part of modelling in the RNA environment.

In general this means that a content item can use every predicate that contains the type of that content item in its domain, and that this content item can link to other content items that are in the ranges of those predicates.

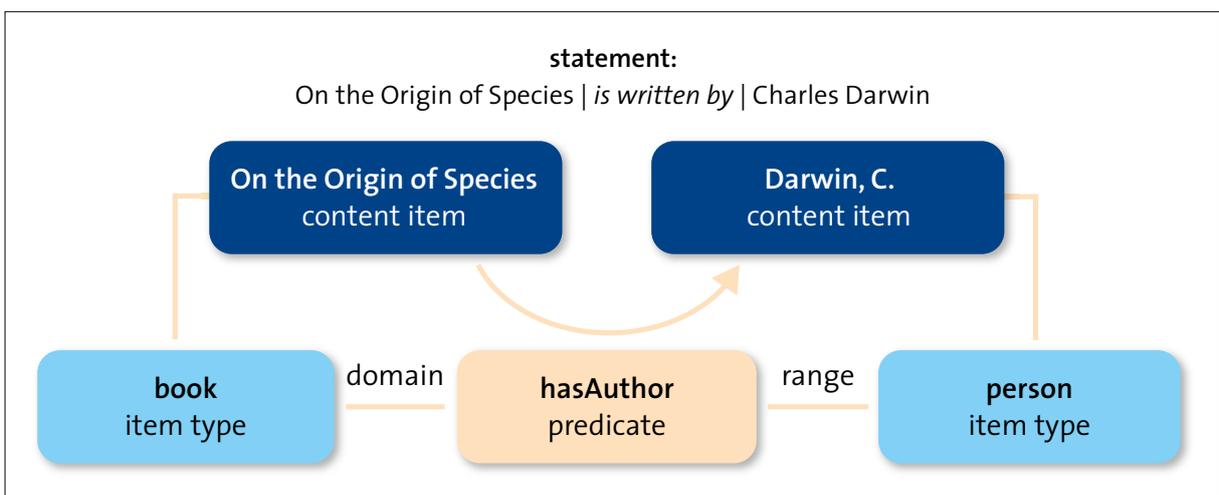


Figure 12: Creating direct relations between content items by using a statement.

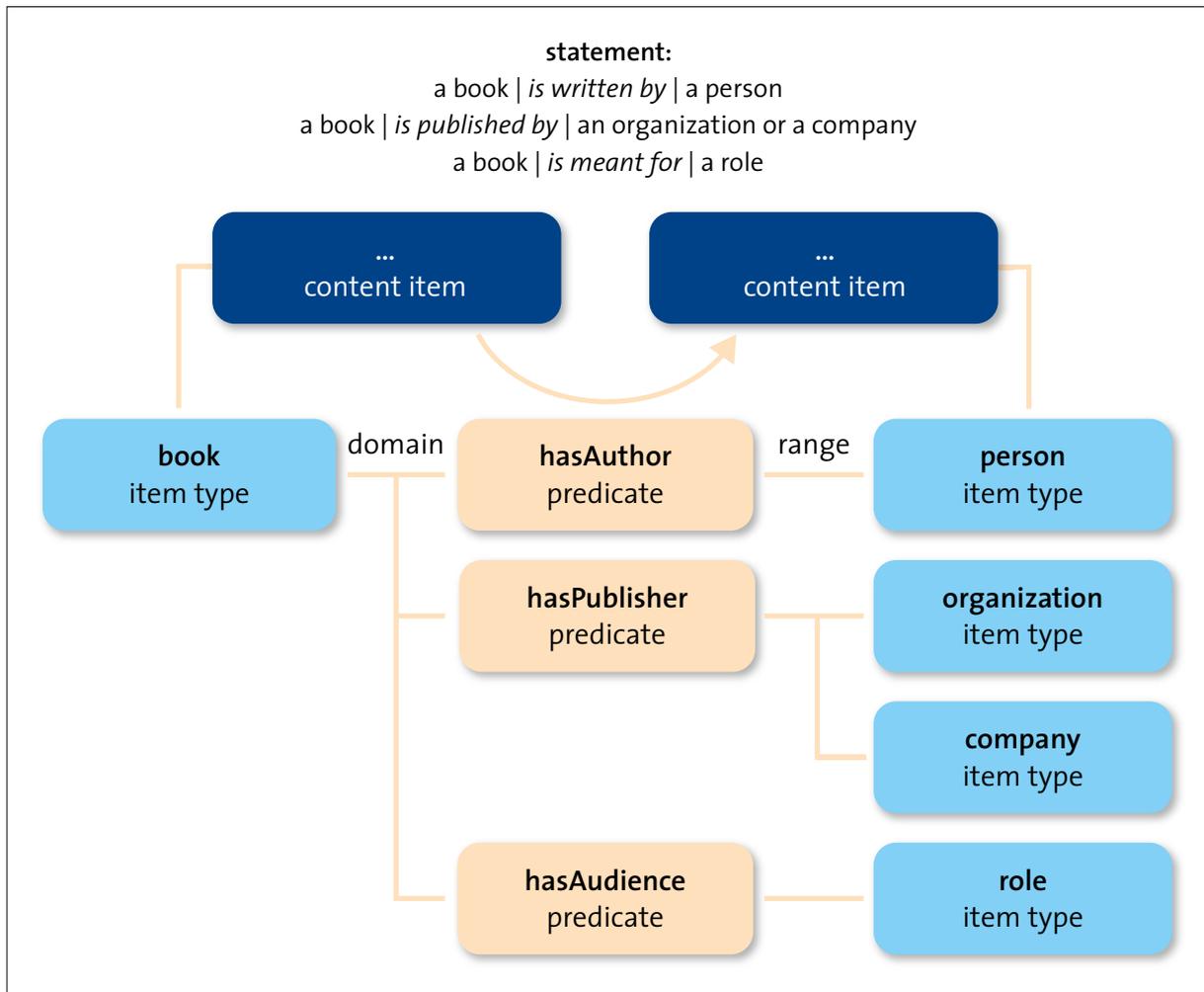


Figure 13: Linking content items within the same domain through a predicate.

In the example above, a content item of the type *book* can link to content items of the type *person*, *organization*, *company* and *role* through the predicates *hasAuthor*, *hasPublisher* and *hasAudience*.

In an RNA environment, predicates and item types can be added when needed. Also item types can be added to domains and ranges of predicates as desired. This enables a highly flexible modelling that can be adapted and extended on the fly.



ANNEX 2: EXAMPLES OF CONTENT ITEMS

Introduction

This annex contains examples of eight types of content items that are frequently used and/or are typical for the STERNA environment. The examples have been derived from the collections of the current STERNA project partners.

The example items all display the three components that every content item is made of:

- “names and description” where preferred names and synonyms of the item are recorded together with a description (all in SKOS format),
- “record data” where the details of the content item are described (metadata) and
- “writing” where rich text can be added to the content item.

Different editors are provided to edit each of these components: a SKOS editor to edit the names and descriptions, an RDF editor for editing record data and a rich-text editor for creating and editing articles in writing.

The “names and description” part supports multilinguality, according to the SKOS standard.

Note that the examples contain several predicates and values that are blue-coloured. These are links to items elsewhere in the RNA environment that represent these predicates and values. Values that are black are not links to an item. These are so-called “literals”; they are nothing more than a string of characters. They can be found with text search, or used for calculations (like numbers or dates), but they are not part of the semantic network.

Also note that several examples display annex items or other content items the item refers to. The RNA Toolset can display these items within the original item itself, without having to navigate from one structure to another.

A third thing to note is that there are several content items of the same item type that seem to represent different types of content. For instance, the third and fourth examples are both of the “specimen” item type, yet the third refers to a mounted specimen and the fourth refers to a drawing. However, both are regarded as specimens. To further specify the type of specimen, a subtype is added in the record data of the items.



Article item type with its annex item *publication event* expanded.

Bibliography of the Megapodiidae

content item / article

names and description [edit | collapse ▲]

preferred label: Bibliography of the Megapodiidae (ENG)

description: The Proceedings of the First International Megapode Symposium (Zoologische Verhandelingen Leiden, vol. 278) published in December 1992 included the paper Bibliography of the Megapodiidae with references to literature on megapodes upto 1992 (Dekker & Jones, 1992: 57-78). The bibliography presented here is supplementary as it continues where the previous list ended. (ENG)

record data [edit | collapse ▲]

item type	article	v
isSubtype	scientific article	v
hasSubjectKeyword	Megapodes	v
hasSubjectKeyword	bibliography	
hasLanguage	English	v
hasAuthor	Dekker, R.W.R.J.	v
hasEvent	publication	☒
	preferred label	publication (ENG)
	item type	publication event
	isPublishedIn	Zoologische Verhandelingen
	hasPublisher	Naturalis
	hasDate	1999
	hasLocation	Leiden
	specification	volume 327
		start page: 169
		end page: 174
hasOwner	Naturalis	v
hasAudience	student	v
hasAudience	scientist	v
hasLink	http://www.repository.naturalis.nl/record/219430	
specification		
note		
source	Naturalis	

Ethnographic artefact item type.

Helmet mask of bwadi dance <i>content item / ethnographic artefact</i>	
names and description [edit collapse ▲]	
preferred label	Helmet mask of bwadi dance (ENG)
preferred label	Masque-heaume de danse bwadi (FRE)
description	Mask of polychrome wood in black and white, provided with tubular eyes and mounted on a crest of feathers. This mask is used in dances, when the dancers put on the figure all the women flee. By tradition, when a masked man has taken away a woman or a child, the proprietor pays a ransom to the kidnapper. The dances of this genre only take place at the end of the rainy season. (ENG)
record data [edit collapse ▲]	
item type	ethnographic artefact v
isSubtype	ethnographic artefact v
isMadeOf	feathers
isMadeOf	wood
isMadeOf	plant fibers
hasEvent	gathering v
hasDonator	Muller, D.
hasRightsStatement	© RMCA, Tervuren, photo: Beaulieux, D.
hasAudience	general audience v
hasLink	http://193.190.223.48/STERNImages/ZoomifyBrowser.php?image=EO.0.0.2101-1_DIA_01
specification	size: 568 x 282 mm; length with fibers: 310 mm Ethnic group: Songye Acquisition type: donation
note	
source	Royal Museum for Central Africa Cultural Anthropology
id in source	EO.0.0.2101-1
writing [edit collapse ▲]	



Specimen item type with the *taxon concept* it refers to expanded.

Τριδάχτυλος Δρυοκολάπτης, ♀

content item / specimen

names and description [edit | collapse ▲]

preferred label Τριδάχτυλος Δρυοκολάπτης, ♀ (GRE)

record data [edit | collapse ▲]

item type specimen ▼

subtype mounted specimen ▼

taxon concept Eurasian Three-toed Woodpecker ⓘ

preferred label Picoides tridactylus (Linnaeus, 1758) (SCI)

preferred label Eurasian Three-toed Woodpecker (ENG)

preferred label Three-toed Woodpecker (ENG)

preferred label pic tridactyle (FRE)

alternative label Picoides tridactylus (SCI)

item type taxon concept

sex female ▼

owner Natural History Museum / Municipality of Amaraoussion ▼

hasExpert Natural History Museum / Municipality of Amaraoussion ▼

audience child ▼

audience scientist ▼

audience student ▼

image [http://www.mfida.gr/\(S\(fhw01m45ikk2utyhyhxiwx45\)\)/Multimedia/MultiPaths/multi_9353.jpg](http://www.mfida.gr/(S(fhw01m45ikk2utyhyhxiwx45))/Multimedia/MultiPaths/multi_9353.jpg)

specification

note Corelation for taxon name ITIS, habitat data Hellenic Ornithological Society

source Natural History Museum / Municipality of Amaraoussion

id in source

writing [edit | collapse ▲]

Specimen item type with annex item *creation event* expanded.

Common Crossbill
content item / specimen

names and description [edit | collapse ▲]

preferred label	Common Crossbill (ENG)
preferred label	Loxia curvirostra (SCI)
description	Illustration of Loxia curvirostra (ENG)

record data [edit | collapse ▲]

item type	specimen	v
isSubtype	specimen drawing	v
taxon concept	Red Crossbill	v
event	creation	🔗
	preferred label	creation (ENG)
	item type	creation event
	creator	Perkins, R.
	date	2009
owner	Archipelagos	v
rights	© R.Perkins / Archipelagos 2009	
audience	general audience	v
audience	student	v
audience	scientist	v
link	http://wildlife-archipelago.gr/birds/common-crossbill/	
specification		
note		
source	Archipelagos	
id in source		

writing [edit | collapse ▲]



Specimen item type.

Sharpe's Rail 87485	
<i>content item / specimen</i>	
names and description [edit collapse ▲]	
preferred label	Büttikofer-ral 87485 (DUT)
preferred label	Sharpe's Rail 87485 (ENG)
record data [edit collapse ▲]	
item type	specimen v
subtype	specimen rotating photograph v
taxon concept	Gallirallus sharpei (Buttikofer, 1893) v
kind of unit	mounted specimen v
age	adult v
event	acquisition event v
event	gathering event v
event	creation event v
owner	Naturalis v
hasExpert	Dekker, R.W.R.J. v
audience	general audience v
link	http://nlbif.eti.uva.nl/naturalis/movies/87485-A.mov
link	http://nlbif.eti.uva.nl/naturalis/movies/87485-B.mov
link	http://nlbif.eti.uva.nl/naturalis/movies/87485-C.mov
rights	© 2005 Naturalis
specification	
note	Holotype Stictolimnas sharpei Büttikofer, 1893. ; Holotype Stictolimnas sharpei Büttikofer, 1893.
source	Naturalis
id in source	RMNH 87485
writing [edit collapse ▲]	



Specimen item type with the content it links to.

Sharpe's Rail 87485
content item / specimen

names and description [edit | collapse ▾]

preferred label Buttikofer-rai 87485 (DUT)

preferred label Sharpe's Rail 87485 (ENG)

record data [edit | collapse ▾]

item type specimen ▾

subtype specimen rotating photograph ▾

taxon concept Gallirallus sharpei (Buttikofer, 1893) ▾

kind of unit mounted specimen ▾

age adult ▾

event acquisition event ▾

event gathering event ▾

event ▾

owner ▾

hasExpert ▾

audience ▾

link

link

link

rights

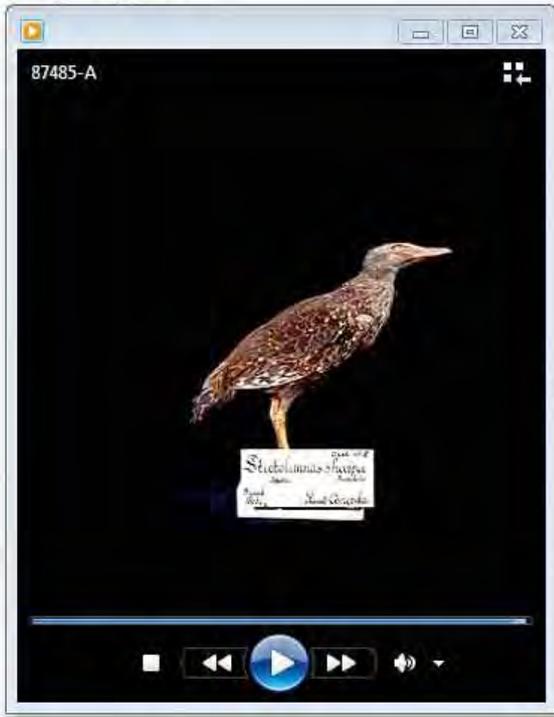
specification

note

source

id in source

writing [edit | collapse ▾]



Stictolimnas

Specimen item type with its annex item *gathering event* folded out.

XC1012 Golden-collared Manakin song

content item / specimen

names and description [edit | collapse ▲]

preferred label XC1012 Golden-collared Manakin song (ENG)

record data [edit | collapse ▲]

item type specimen v

subtype bird sound v

taxon concept Golden-collared Manakin v

event gathering [expand]

preferred label gathering (ENG)

item type gathering event

gathering agent Carter, R.

date 2001-05-25

date specification ?

location Panama

location specification El Valle de Anton, Coclé

longitude -80.125100000000003

latitude 8.6000999999999994

altitude 600 m

rights Copyright on this recording lies with the recordist; it is published on xeno-canto.org under a Creative Commons Attribution-NonCommerical-NoDerivs 2.5 License

link <http://www.xeno-canto.org/1012>

specification buzz

note

source Xeno-canto

id in source XC1012

writing [edit | collapse ▲]



Taxon description item type.

Major Mitchell's Cockatoo

content item / taxon description

names and description [edit | collapse ▲]

preferred label	Cacatua leadbeateri leadbeateri (Vigors, 1831) (SCI)
preferred label	Major Mitchell's Cockatoo (ENG)
preferred label	Cacatoès de Leadbeater (FRE)
preferred label	Inkakakadu (GER)

record data [edit | collapse ▲]

item type	taxon description	v
subtype	taxon webpage	v
taxon concept	Cacatua leadbeateri leadbeateri (Vigors, 1831)	
language	French	v
owner	Natural History Museum of Luxembourg	v
rights	© Musée national d'Histoire naturelle Luxembourg	
audience	bird watcher	v
audience	scientist	v
audience	general audience	v

specification

note

source Natural History Museum of Luxembourg

id in source

writing [edit | collapse ▲]

Major Mitchell's Cockatoo

Cacatua l. leadbeateri (Vigors, 1831)



- Taille : 35 cm.
- Poids : 300-435 gr.
- D: Inkakakadu F: Cacatoès de Leadbeater

Taxon presence item type.

Melodious Warbler

content item / taxon presence

names and description [edit | collapse ▾]

preferred label	kratkoperuti vrtnik (SLV)
preferred label	Melodious Warbler (ENG)

record data [edit | collapse ▾]

item type	taxon presence	v
subtype	taxon presence map	v
hasTaxonConcept	Melodious Warbler	v
link	http://www.gis.si/noags/dopps_pregled.php?vrsta=292	
specification		
note	Podatki še niso zbrani. To so delni rezultati.	
source	DOPPS	
id in source		

writing [edit | collapse ▾]



ANNEX 3: KNOWING YOUR WAY AROUND THE RNA TOOLSET

A: anatomy of the RNA Toolset

B: most-used operations in the RNA Toolset

Creating a new structure

Opening a structure

Creating a new item

Editing a content item

Adding an annex item to a content item

Adjusting view settings



A: ANATOMY OF THE RNA TOOLSET

The RNA Toolset has been developed by Trezorix and is made available to the content providing partners online as Software as a Service (SaaS).

Accessing the RNA Toolset

To use the RNA Toolset, all you need in technical terms is a computer with access to the Internet. You can enter the RNA Toolset at:

<http://rnatoolset.sterna-project.eu>

On the RNA Toolset access page you will be asked to enter your (user-) name and password. A system administrator, who has the authorization to create new user profiles, provides you with this login data. If you do not have a user name and password or if you have lost your user name and/or password, please contact a system administrator at 'sterna AT naturalis.nl'.

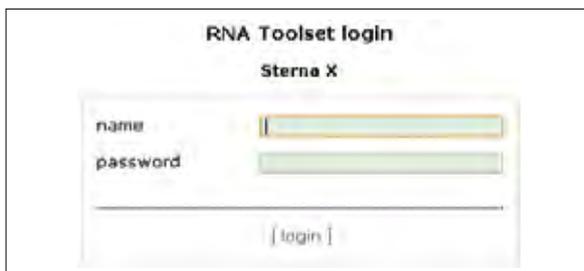


Figure 14: RNA Toolset login

To proceed, enter your name and password and click on the [login] button.

Overview screen

After you have logged in, the RNA Toolset displays the *overview* screen. This screen consists of three panels. In the left panel, all the structures that are present in the STERNA RNA environment are displayed, together with several basic functionalities, such as the options to create new structures and folders, move structures and folders and access the view settings dialogue, which is explained in part B of this annex.

In the overview screen, the centre panel is empty, except for the [management] button at the bottom of the screen. This gives you access to the *management* menu. The various functions of this menu are explained in part B of this annex.

The right panel contains the search functions of the RNA Toolset. You can use these to search for content in the STERNA RNA environment with full-text searches and facet searches. The right panel also contains a tracking function, which allows you to track certain actions performed by yourself or by other partners on the content items that are stored within the RNA environment. This might for instance be very useful when a specific workflow for the content enrichment procedure has

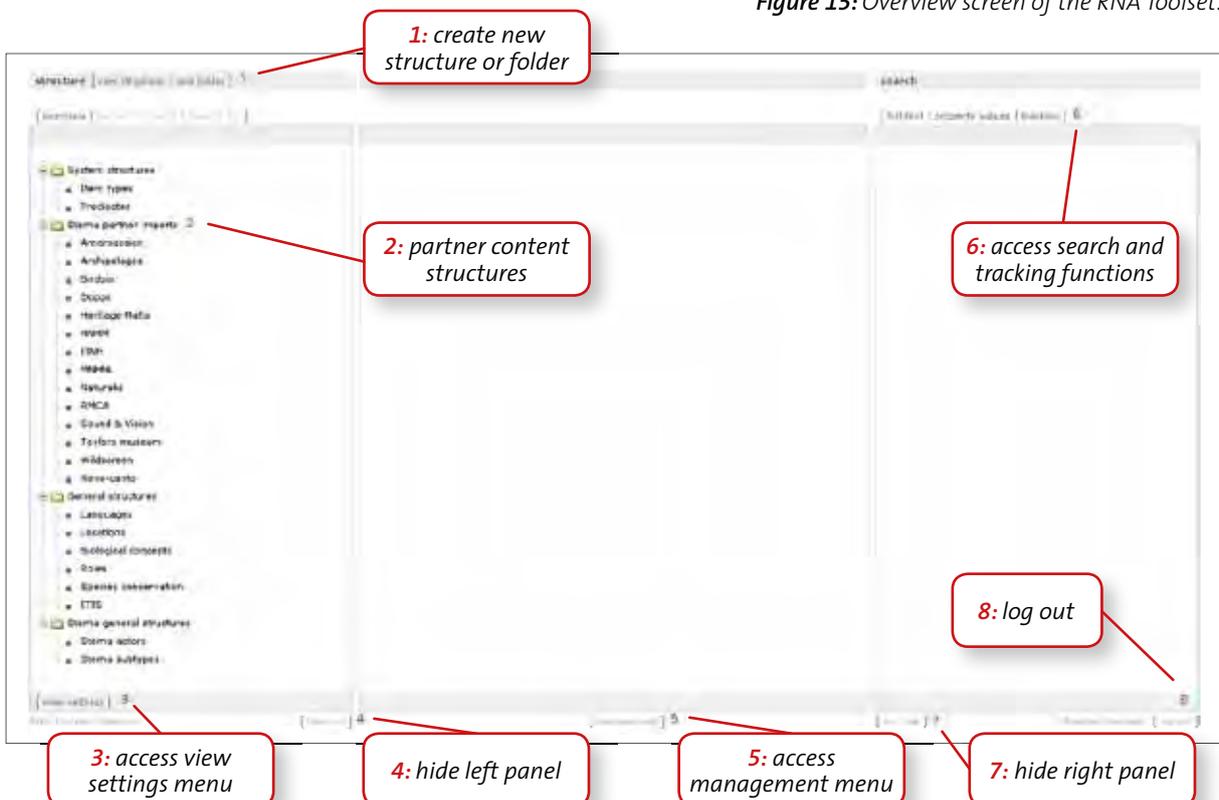


Figure 15: Overview screen of the RNA Toolset.

been designed (see paragraph 6.1). Search results are also displayed in the right panel.

The screenshot below displays the overview screen of the RNA toolset.

Note that there might be differences between the overview screen displayed above and the actual overview screen displayed when you have logged on to the RNA Toolset, with respect to the available buttons. Which buttons are and are not available depends on the authorization scheme of your personal user profile.

Current structure screen

When you open one of the structures in the left panel, the *current structure* screen is displayed. The left panel now displays the chosen structure, which can be unfolded further by clicking on the **[+]** buttons in front of the displayed items. Also several new buttons have appeared, such as various move buttons, a **[link]** button and a **[structure settings]** button.

When you click on a content item in the structure, its content is displayed in the centre panel. Each content item consists of three components. At the top, the *names and description* of the content item is displayed. Here, preferred names in several languages and synonyms of the item are recorded, along with a description of the

item, possibly also in different languages. Directly below, the item's *record data* is displayed. This section contains all the (meta-) data that describes the item and its context. At the bottom there is a section called *writing*. Here, rich text objects can be recorded, such as web articles with images.

Each of these components, *names and description*, *record data* and *writing* has its own editor. The names and description part of the item can be edited with the SKOS editor, while the RDF editor is used to edit the record data. Writing can be edited with the use of the rich text editor.

In addition to the content item, you will also notice various new functions in the centre panel, such as the buttons to create a new item or annex item and to delete a selected item. These will be explained in part B of this annex. Also available are the tab pages:

- **[referrers]** which shows what other content items refer to the current item,
- **[system]** where system information such as the item's identifier is displayed,
- **[tracking]** which shows who has done what with the selected content item, and
- **[comments]** where editorial comments regarding the selected item can be entered and reviewed.

The right panel in the current structure panel is the same as in the overview panel and contains the search

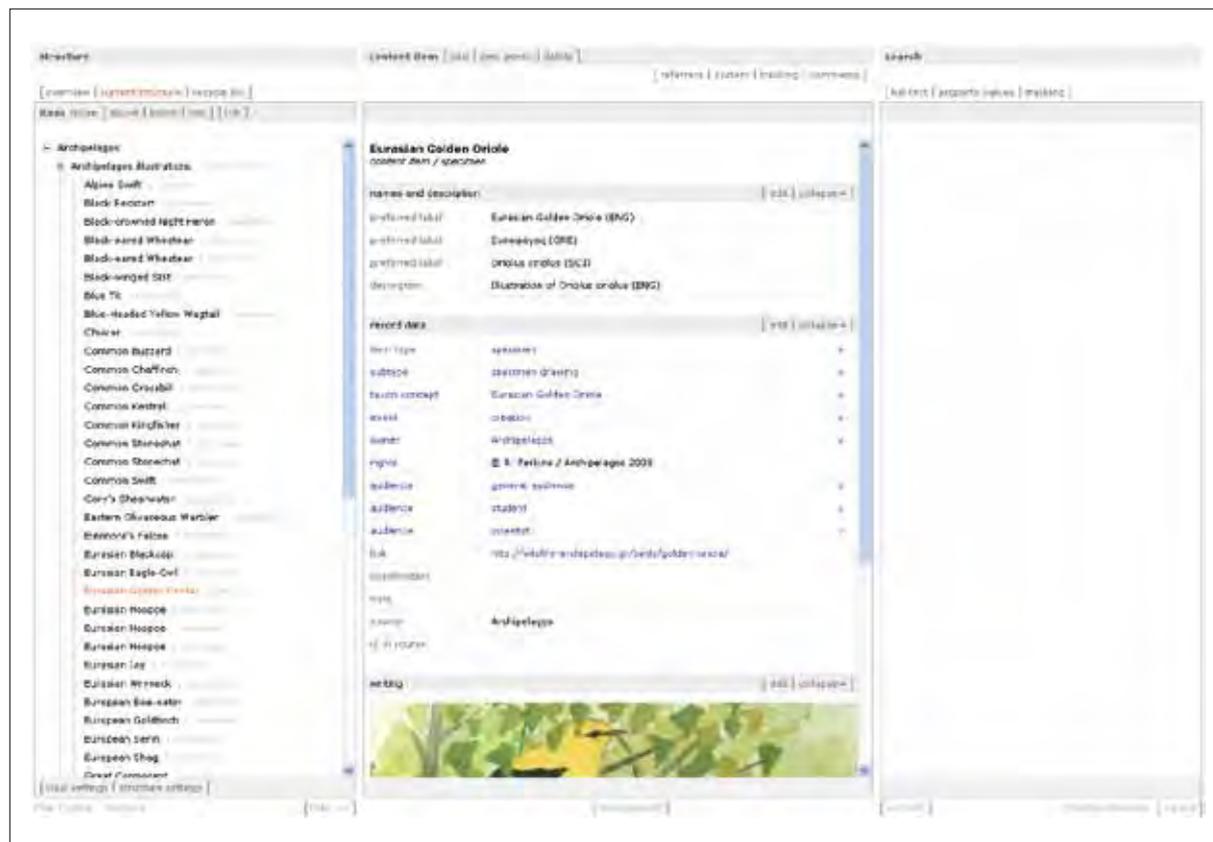


Figure 16: Structure panel and editing panel of the RNA Toolset.

and tracking functions of the RNA Toolset. The figure below displays a screenshot of the current structure panel with a content item in the centre panel.

Different roles for different users

Not every user can perform the same actions in the RNA Toolset. Which actions you can and cannot perform depends on two aspects: the authorizations assigned to your user profile and the edit permissions within each structure.

With regard to user authorizations, there are five general actions that a user either is or is not allowed to perform. These are:

- Manage reference structures: this allows users to create and edit reference structures
- Manage authorization: this enables users to determine the general actions that other users are allowed to perform
- Manage users and groups: having this authorization allows users to create and delete user accounts and to create, manage and delete groups of users
- Manage imports: this enables users to operate the Excel Generic connector
- Value mapping: this enables users to create multiple property links between content items in a fast and semi-automated way (see part C of this annex)

Depending on the role of a user, different combinations of these authorizations can be made.

With regard to structure settings, each structure has its own settings with respect to which user or group may perform which action regarding the content items within the structure.



B: MOST-USED OPERATIONS IN THE RNA TOOLSET

In part A of this annex we introduced you to the basics of the RNA Toolset. In this part, we describe some of the operations that are performed most commonly when working with the RNA Toolset. These are:

- creating a new structure
- opening a structure
- creating a new item
- editing a content item
- adding a new annex item to a content item
- adjusting view settings

The complete RNA Toolset manual is available online at:
<http://manual.rnaviewer.net>

Creating a new structure

Provided that you are authorized to create and manage reference structures, the RNA Toolset provides you with an easy method to create new structures. These may contain content items and *simple nodes*, items that only serve as a means to hierarchically order a structure. Also they may contain item names and descriptions in any language you choose. This is how a new structure is created:

- In the upper left corner of the *overview* screen, click on the [new structure] button (see figure 15).

- In the dialogue that appears, first enter the name of the structure in the text box below “title”. The name of the structure should be representative of the content it will contain.
- Choose the languages in which the structure’s content items can have labels:
 - o Select a language from the left list below “select allowed languages” by clicking on it (select several languages simultaneously by holding the Ctrl key while clicking on the languages).
 - o Click [>>] to transfer the selected language to the list on the right (allowed).
- Set the structure’s preferred language, the language that is attributed automatically to the first label of new items in the structure:
 - o In the list on the right (allowed), click on the preferred language.
- Click on move item [up] until this language is at the top of the list.
- Below “select allowed content types”, choose the base types (see annex 1) the structure may contain. For a standard structure containing content, always choose “content item” (“simple node” is always selected automatically, as this type of item may occur in every kind of structure).
- Click on [ok].

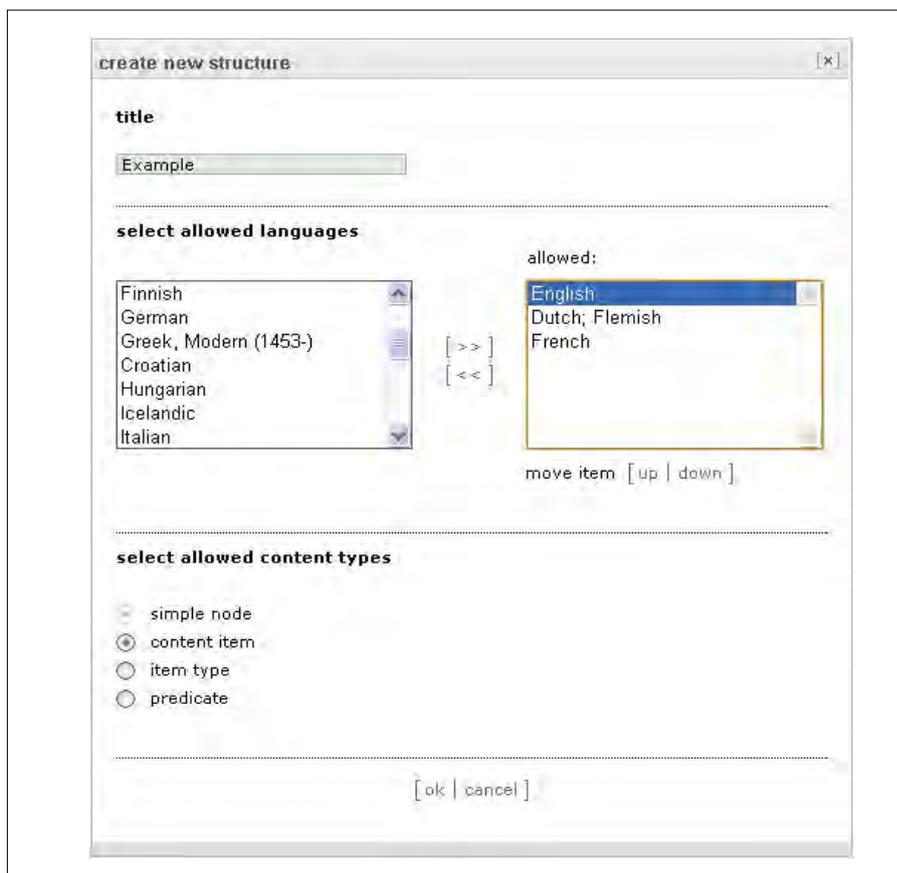


Figure 17: Dialogue box for creating new structures.

If others want to use the newly created reference structure, they need to get permission for the various actions that can be performed on a structure. They should consult the system administrator to obtain these rights.

Opening a structure

The structures in the left panel of the *overview* screen can be opened as follows:

- Select the structure that is to be opened by single-clicking on it. The structure name turns red.
- Click on **structure [open]** at the top of the left panel.

The RNA Toolset now opens the *current structure* screen. The left panel displays the open structure.

Now the structure can be folded out by clicking on the **[+]** buttons to the left of the content items. Click on the **[-]** buttons left of the content to collapse (parts of) the structure again. Content items that do not have a **[+]** button are on the deepest level.

As an alternative, a structure in overview mode can also be opened by double-clicking on it.

Creating a new item

Creating new items will be one of the most performed actions when enriching your content manually. Here is how it works.

- Open the structure you want to add a new item to.
- In the structure, navigate to the item in which the new item will be placed below or into.
- Select this content item by single-clicking on it. The name of the item turns red.
- Click on **content item [new]** at the top of the centre panel. A dialogue will now appear.

In the dialogue box shown below, first choose the base type. Choose between:

- **simple node:** this base type only serves as a means to hierarchically order a structure and can only have names and a description (SKOS-only).
- **content item:** this base type can be used to create and edit content and consists of three components, a *names and description* part, a *record data* part and a *writing* part.

Second, choose the exact place of the new item in the structure. Choose between:

- **Create item as child of current item:** the new item will be placed as a “child” item one level below the selected item.
- **Create item on same level as current item:** the new item will be placed with the selected item on the same level.

When you have chosen a content item, you also need to determine what type of content item you want. The predicates (properties) that are available in the *record data* part of your item depend on the type of content item you choose (see annex 1).

- Click on **[select type]**.
- In the list of item types, navigate to the item type of your choice.
- Select it by single-clicking on it. The name of the item type turns red.
- Click on **[ok]**.

In the text box called “title” enter the name of the new content item. This should be in the structure’s preferred language.

Click on **[ok]**.

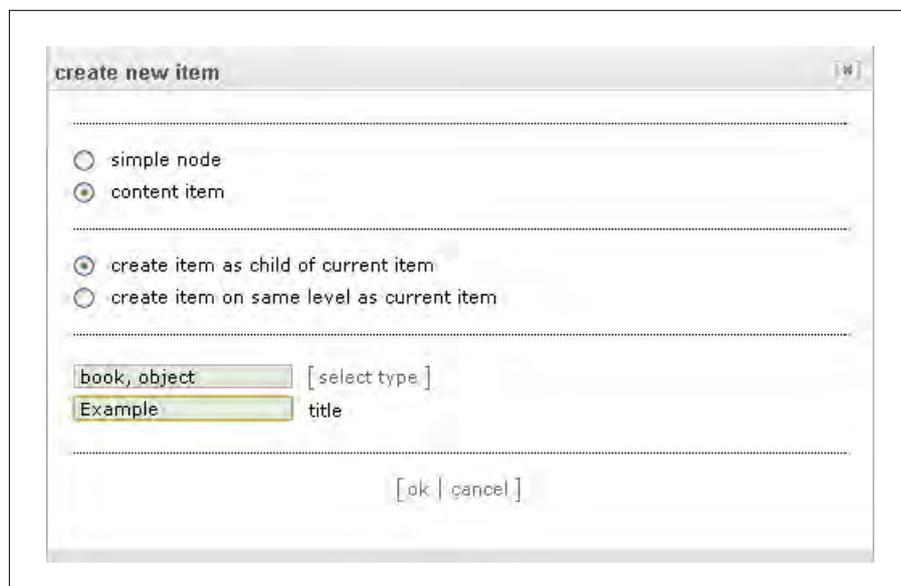


Figure 18: Creating new items – dialogue box.

Editing a content item

Through editing your content item, you can add names and definitions in different languages, metadata and links to other items as well as rich text. Some of the basic editing skills are explained here.

Open the right editor

All components of a content item, *names and description*, *record data* and *writing* have their own editor. To open the editor of the component you want to edit, simply click on its **[edit]** button.

Eurasian Golden Oriole
content item / specimen

names and description [edit | collapse ▾]

preferred label	Eurasian Golden Oriole (ENG)
preferred label	Συκοφάγος (GRE)
preferred label	Oriolus oriolus (SCI)
description	Illustration of Oriolus oriolus (ENG)

record data [edit | collapse ▾]

item type	specimen	v
subtype	specimen drawing	v
taxon concept	Eurasian Golden Oriole	v
event	creation	v
owner	Archipelagos	v
rights	© R. Perkins / Archipelagos 2008	
audience	general audience	v
audience	student	v
audience	scientist	v
link	http://wildlife-archipelago.gr/birds/golden-oriole/	
specification		
note		
source	Archipelagos	
id in source		

writing [edit | collapse ▾]

Figure 19: Opening an editor to edit reference structures and content items.

Editing names and description

The editor for the names and description part of the content item is a SKOS editor. This means that you can add and edit information about the content item that is in accordance with the SKOS standard, a semantic web standard for knowledge systems. The SKOS editor enables you to add preferred and alternative names to the content item as well as descriptions of the item in several different languages.

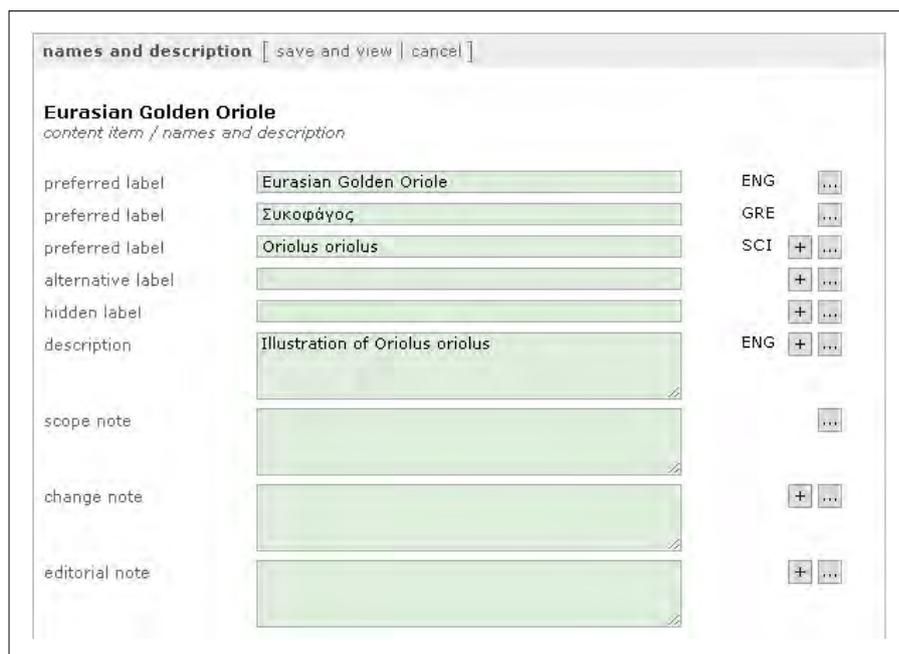


Figure 20: Adding new preferred labels with the SKOS editor.

As you can see in figure 20 above, next to all of the labels and notes (except for the scope note) there is a **[+]** button. This means that the label or note is repeatable, i.e. that it may contain more than one value. Clicking on the **[+]** button will add a second text box for the same label or note. There is no limit to the number of repeated labels for a label or note type. However, a content item can have only one preferred label in each available language.

Also note that there is a **[...]** button next to all of the labels and notes. By clicking on this you open a menu that helps you to attribute the right language to the label or note (see figure 21). This menu has some additional features, such as the possibility to change the type of the label, to move it and to delete it.



Figure 21: Selecting the right languages for labels and notes.

To set the language of a label or note, open the dropdown menu “language” and select the right language from the list by clicking on it.

After editing the names and description of the content item, click on **[save and view]** to store your changes.

Editing record data

The editor for *record data* is an RDF editor. This editor enables you to add and edit information that gives further details about a content item, such as its subject keyword(s), its related taxon concept(s), its creator, when and how it was made and who owns it. This metadata is stored in RDF format, the semantic web standard for meta- and context data.

Figure 22 shows the record data editor with an empty record for a *specimen* content item type, called “Mounted specimen of White stork”. Note that all of the available fields, except the “item type” field, have a **[+]** and a **[-]** button next to them. This means that all of these fields can be repeated (by clicking on the **[+]** button) and that a repeated field can be deleted again (by clicking on the **[-]** button).

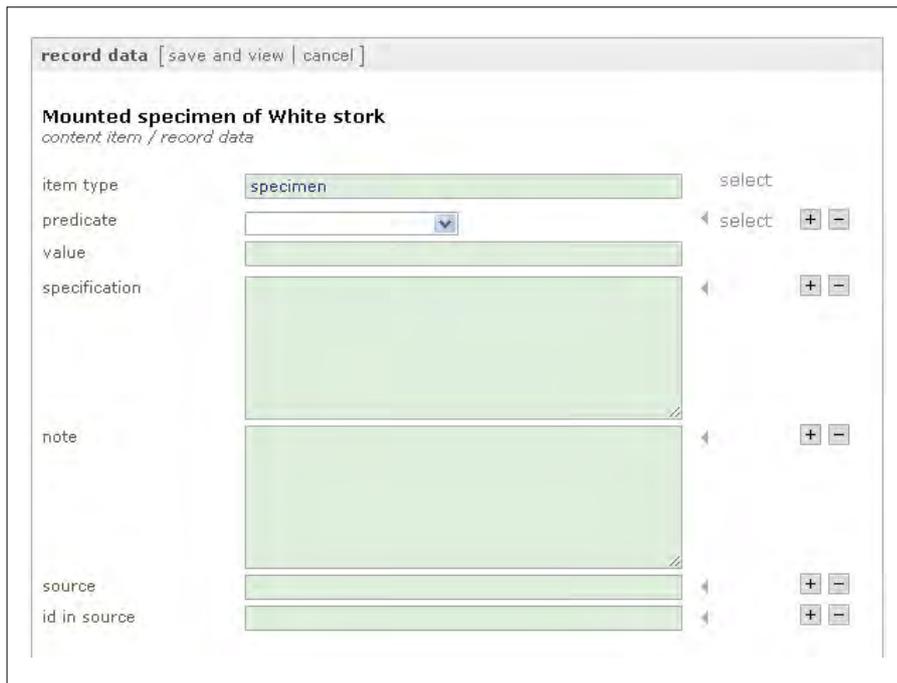


Figure 22: Record data editor of the RNA Toolset.

Also note that there is a **select** button next to two fields. This can be used to select a value for these fields from one or more structures in the STERNA RNA environment, thus creating a semantic link between this content item and the item in another structure that represents the value for the field.

A third important thing to note here is the drop-down menu next to “predicate” and the “value” field below it (see figure 23). The drop-down list can be used to select predicates (see annex 1). The list contains all the predicates that have the selected item type in their domain. The “value field” can either contain a literal or (preferably) a semantic link to another content item.

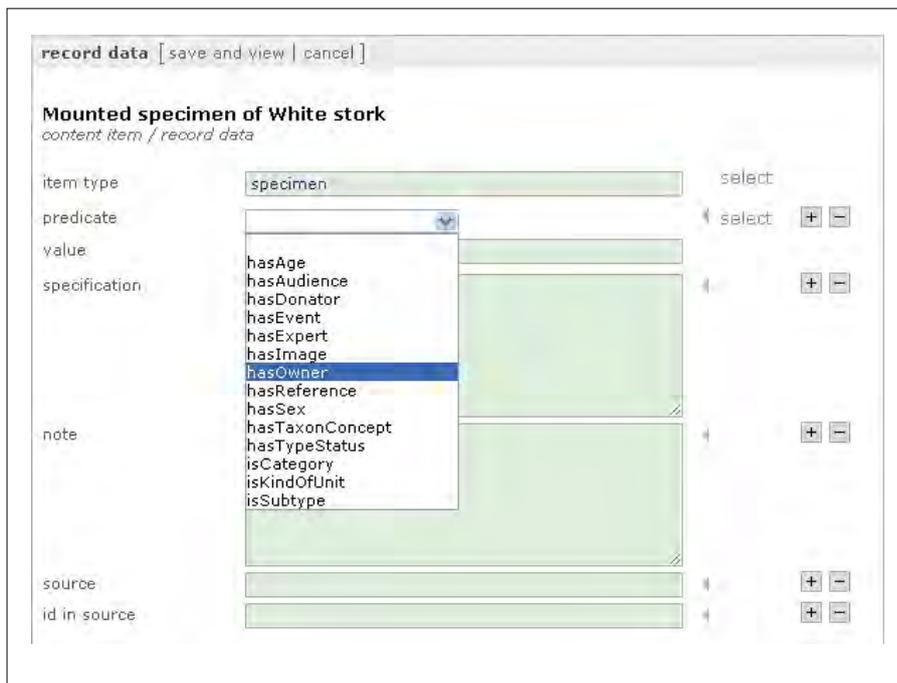


Figure 23: Selecting predicates from a drop-down list.

record data [save and view | cancel]

Mounted specimen of White stork
content item / record data

item type: specimen [select]

predicate: isSubtype [+ -]

value: mounted specimen

predicate: hasTaxonConcept [-]

value: Ciconia ciconia (Linnaeus, 1758)

predicate: hasAge [-]

value: adult

predicate: hasSex [-]

value: male

predicate: hasOwner [-]

value: Naturalis

predicate: hasAudience [-]

value: scientist

predicate: hasImage [-] [select]

value: http://www.example.org/white-stork.jpg

specification: height: 125 cm [+ -]

note: gathering information incomplete [+ -]

source: Naturalis [+ -]

id in source: RMNH 123.456 [+ -]

Figure 24: Available predicates associated with a content item type.

Which predicates are available depends on the chosen content item type. To every content item type a set of predicates is attributed that can be used to describe the characteristics that are typical for that type (see figure 24).

Editing writing

Finally, with the rich text editor you can create and edit rich text items within the content item itself, for instance to create web articles with images. As shown below, the rich text editor is easy to use and contains most of the standard functions of a basic web content editor.

writing [save and view | cancel]

Rich text editor toolbar with icons for Bold (B), Italic (I), Underline (U), Paragraph, Undo, Redo, Cut, Copy, Paste, Link, Unlink, Table, HTML, Print, and other standard editing functions.

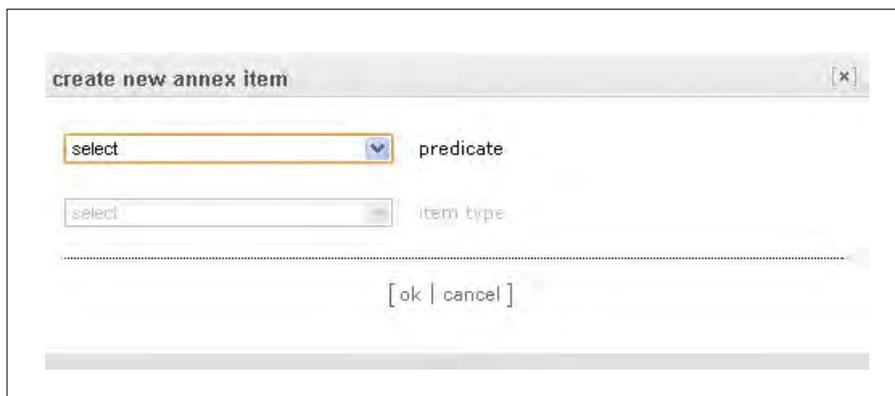
Figure 25: Interface of the rich text editor.

Adding an annex item to a content item

Content items can have one or more **annex items**. These are items that are closely coupled to another item, their parent item. They do not have much meaning on their own, as their meaning comes from their parent item. When in a search you find a property in an annex item, it will point to the parent item, not to the annex item. And an annex item does not show in the reference structure like other items do. Typical examples of the use of annex items are *events* (a *creation event* only has meaning in combination with its parent, the object that is created). When you want to add an annex item to a content item, this is how it works:

- Select the content item you want to add an annex item to by single-clicking on it.
- Click on **[new annex]** at the top of the centre panel.

The following dialogue appears:



- First select the predicate you want to use to link the parent item and the new annex item. This can be selected from the “predicate” drop-down list. When your annex item is for instance a gathering event, select the predicate “hasEvent” from the list.
- After selecting the predicate, the “item type” drop-down menu becomes active. From this list, select the type of content item of your new annex item, in this case the item type “gathering event”.
- Click on **[ok]**.

Now the annex item can be edited like any regular content item.



Figure 26: Adding an annex item.

When you open the annex item's parent item, you will see that it now has a predicate "hasEvent" with a link to the annex item as value. To view this annex item within its parent item, click on the **V** right of the link as depicted below. To navigate to the annex item itself (so you will be able to edit it), click on its name. Figure 27 shows how to open an annex item.

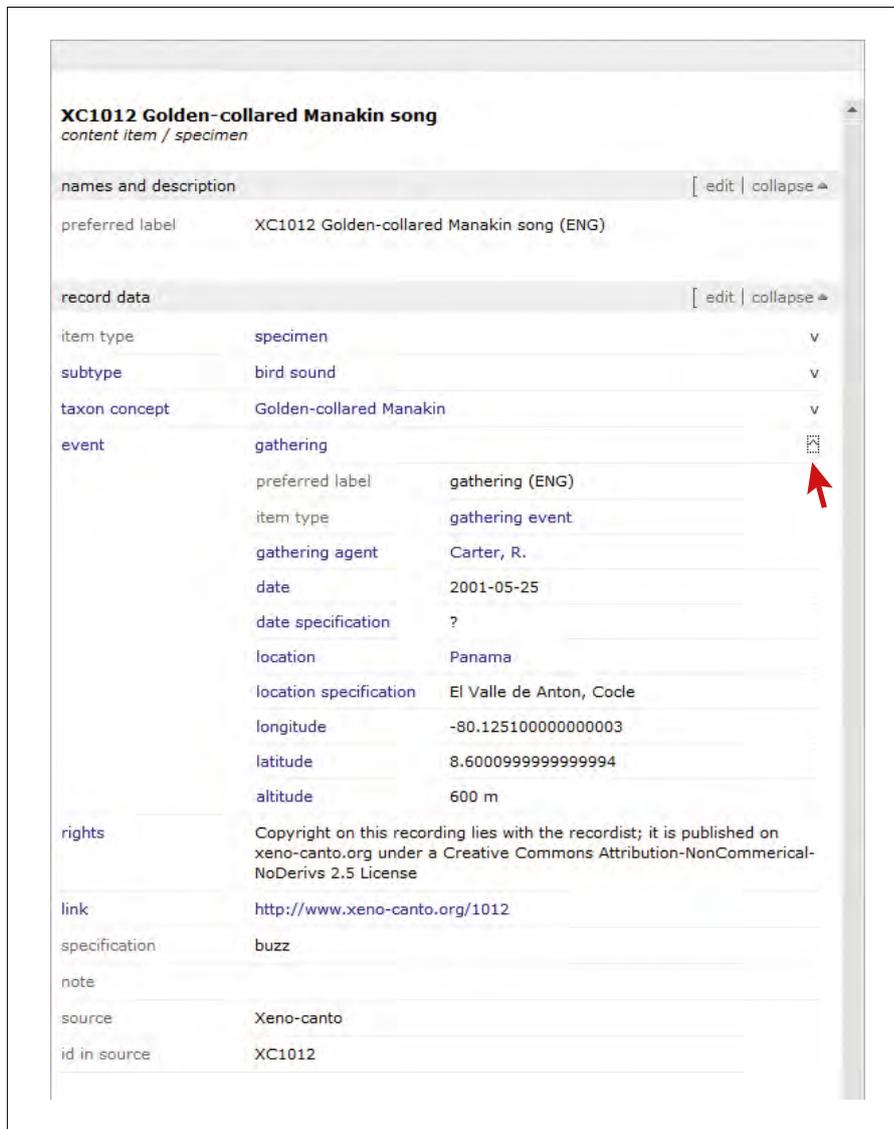


Figure 27: Opening an annex item by clicking on "V".

Adjusting view settings

In the "view settings" menu (figure 28) you can adjust several settings with respect to the way structures and items are displayed to your own preferences. The "view settings" menu can be accessed by clicking on the **[view settings]** button in the lower left corner of the **overview** screen or the **current structure** screen.

The settings you can adjust in the "view settings" menu are:

- **[save]** and **[view]** buttons
- the presentation of item labels with or without their item type
- the maximum number of paginated items
- the language in which the names of the items in structures are displayed in

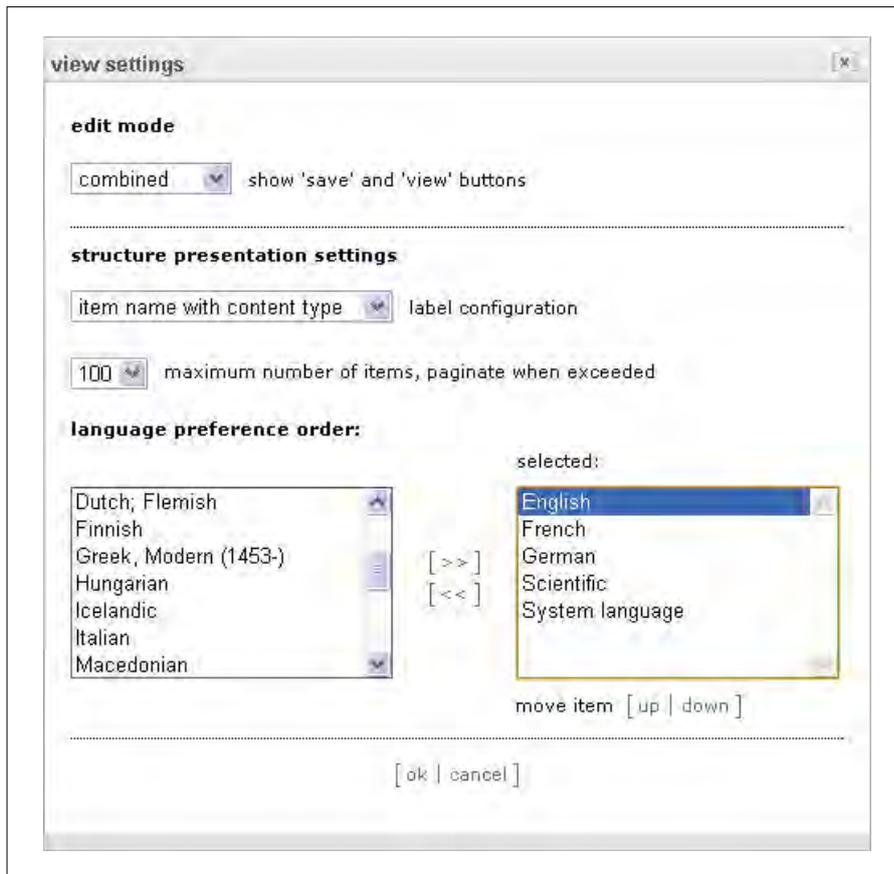


Figure 28: Adjusting the view settings.

With the “show ‘save’ and ‘view’ buttons” drop-down menu you can choose whether you want the **[save]** and the **[view]** buttons that are at your disposal when you are using one of the editors to be presented as two separate buttons or as one **[save and view]** button. The first option will leave the item you are editing in edit mode after saving your changes, the last option will change it to view mode automatically after saving.

With the “label configuration” drop-down menu you can change the way items are presented in their structure. You can choose between *item name only* and *item name with content type*. The first option will only show the item name, the second one adds the item type to the item name.

In a structure, the content items below every node are being paginated (divided in “pages”) to enable users to navigate through them more efficiently. The number of content items per “page” can be altered with the drop-down menu called “maximum number of items, paginate when exceeded”.

With the help of the language lists below “language preference order” you can decide in which language you want the names of items in a structure to be presented. If you would like your names to be presented for instance in French, click on French in the selected box and move French to the top of the languages list by clicking on the **[up]** button several times until French is at the top of the list of selected languages.

When done adjusting your view settings, click on **[ok]**. For the new settings to take effect, click on the **[overview]** button or the **[current structure]** button (depending on which screen is active, overview or current structure) or press **Ctrl + F5**.



GLOSSARY

annex item

Content items that do not have much meaning as independent items, but only serve as a container for a certain set of properties of the content item they belong to, the *parent item*. Examples of annex items are *events*. Annex items are incorporated more or less inside their parent item, which has certain advantages for findability.

base type

In the RNA environment there are four base types: *content item*, *item type*, *predicate* and *simple node*.

child item

Item that is placed hierarchically below another item, it is the lower part of a parent – *child* relation.

content

Textual, visual and audio information units that can be approached through the web. This can either be *unstructured* content, like articles, or *structured* content, like database records.

content item

Items directly related to content. They consist of three building blocks, *names and description* (SKOS data), *record data* (mainly RDF metadata) and *writing* (rich text data in XML format).

item

Identifiable unit of information. Items and the links between them are the building blocks of the RNA environment.

item type

Item types are used to indicate the type of a content item (e.g. article, video, language, taxon concept, etc.).

metadata

Literally data about data. Metadata is used to describe all kinds of details about content, such as its creator, subject, owner, property rights etc.

parent item

Item that is placed hierarchically above another item, it is the upper part of a *parent* – *child* relation.

predicate

Predicates typify the relations (i.e. the links) between two items, in a statement of the form *subject* – *predicate* – *value* (*object*). Such a statement is called a *property*.

RDF

Resource Description Framework, a standard that has been developed for modelling web content, especially to improve the findability of web content. RDF plays a key role in the semantic web, also known as Web 3.0.

reference network

A network of content items connected through references in their properties to other items.

rich text

A format for creating media-enriched content, such as web articles with images.

RNA

Reference Network Architecture, an architecture for reference networks. The RNA architecture is mainly based on the standards XML, RDF and SKOS.

RNA Toolset

The RNA Toolset is a set of tools that allows users to create and edit data in the RDF layer of an RNA environment. With this toolset, users can create and edit content items.

simple node

An item that contains only SKOS data (names and description) and is mainly used for ordering purposes in reference structures.

SKOS

Simple Knowledge Organization System, a modern, web-oriented version of the 1986 ISO 2788 standard for thesauri. SKOS is based on RDF.

XML

Extensible Markup Language, a W3C standard for structuring web content.

FOOTNOTES

- 1 For the current STERNA project, we defined four initial use cases focussing on birdwatchers, a digitally savvy, young audience, people at sea (be it tourists or boaters), and 'humble rambblers' (hikers). A fifth use case focuses on selecting STERNA content that may also be of interest to the audiences of the Europeana-portal.
- 2 See <http://dublincore.org/documents/dcmi-terms/#terms-audience>
- 3 From: David Weinberger, *Everything Is Miscellaneous: The Power of the New Digital Disorder* (New York 2007)
- 4 For an overview of file formats see for instance <http://www.fileinfo.net/common.php>
- 5 OAI: Open Archives Initiative – see www.openarchives.org
- 6 TAPIR: TDWG Access Protocol for Information Retrieval – see www.tdwg.org/activities/tapir
- 7 For an explanation on domain and range, see Annex 1: Data modelling in STERNA.
- 8 See DEL 5.3.1 STERNA Evaluation Methodology, available upon request.
- 9 RNA project: a SenterNovem Prima project, 2005-2007, see www.rna-project.org
- 10 The RNA Toolset has been developed and is owned by Trezorix BV, <http://www.trezorix.nl>.



IMPRESSUM

IPR, reuse and citation information

© Copyright of this document: STERNA Consortium and Trezorix BV, 2009-2010.

All information in this document is the property of The STERNA Consortium and Trezorix BV. This document may not be copied or published in any way, either as a whole or in part, without expressly stated consent of The STERNA Consortium and Trezorix BV.

© Copyright RNA Toolset: Trezorix BV, Delft, Netherlands, 2009-2010

The description of the RNA architecture and the RNA Toolset is the property of Trezorix BV. RNA Toolset is a registered trademark of Trezorix BV.

Official citation

Nederbragt, H., Heerlien, M.(2010): Methodology for Content Enrichment.

Disclaimer

This report was produced by the STERNA project with the financial support of the European Commission. The content is the sole responsibility of STERNA and its project partners. Furthermore, the information contained in the report, including any expression of opinion and any projection or forecast, does not necessarily reflect the views of the European Commission and in no way anticipates any future policy plans in the areas addressed in this report. The information supplied herein is without any obligation and should be used with the understanding that any person or legal body who acts upon it or otherwise changes its position in reliance thereon does so entirely at their own risk.

Imprint

Authors:

Hans Nederbragt, Maarten Heerlien (Trezorix)

Reviewers:

Andreas Strasser, Salzburg Research
Andrea Mulrenin, Salzburg Research
Sander Pieterse, NCB Naturalis
Wernher Behrendt, Salzburg Research

Graphics & layout:

Daniela Gnad, Salzburg Research

Images:

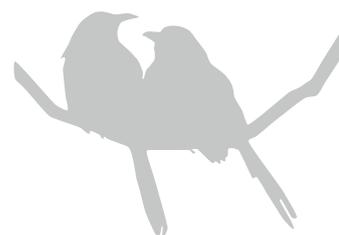
Image on cover and page 1: © stock.xchng
Image on page 15: © Teylers Museum
Bird illustrations: © Shutterstock
Screenshots of RNA Toolset in the annex: © Trezorix BV

Copyright of the report:

Salzburg Research on behalf of the STERNA Consortium

ISBN 978-3-902448-23-1

Printed in Austria, 2010





Project Coordinator

salzburg|research

Salzburg Research Forschungsgesellschaft m.b.H.

Jakob Haringer Straße 5/3 | 5020 Salzburg, Austria

Phone: +43.662.2288-201 | Fax: +43.662.2288-222

Project website: <http://www.sterna-net.eu>

Project Management

Andrea M. Mulrenin

andrea.mulrenin@salzburgresearch.at

The STERNA Consortium and Contributors



The STERNA project is supported and partly funded by the eContentplus programme of the European Commission.

ISBN 978-3-902448-23-1