ENDURE

European Network for Durable Exploitation of crop protection strategies

Project number: 031499

Network of Excellence
Sixth Framework Programme

Thematic Priority 5
FOOD and Quality and Safety

Deliverable DS4.8

Modular concept to extract, validate and transfer data from IA4 to Endure-IC.

Due date of deliverable: M18

Actual submission date: M19

Start date of the project: January 1st, 2007 **Duration**: 48 months

Organisation name of lead contractor: JKI

Revision: V1

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)	
Dissemination Level	
PU Public	
PP Restricted to other programme participants (including the Commission Services)	**X**
RE Restricted to a group specified by the consortium (including the Commission Services)	
CO Confidential, only for members of the consortium (including the Commission Services)	

Table of contents

Glossary

Content	summarizes the data intended for exchange. In the document the content is the reports with all its attachments of Endure ALPS intended for publishing.
CQL	or Contextual Query Language is a query language provided by SRU.
Data	means logically grouped information units
Data Provider	is the one providing content.
EPPO	or European and Mediterranean Plant Protection Organization is an intergovernmental organisation responsible for European cooperation in plant health.
EIC	ENDURE Information Centre
HTTP	or Hypertext Transfer Protocol defines mechanisms for common data transfer via a network
OAI	is the Open Archives Initiative.
OAI-PMH	or OAI Protocol for Metadata Harvesting is a protocol for XML-based data exchange made by OAI.
Recipient	is the one consuming content.
SOAP	is a XML-based protocol describing the interface of a web service.
SRU	or Search / Retrieve via URL is a protocol for XML-based data exchange.
Web Service	describes a common standard for accessing remote systems.
XML	or Extensible Markup Language provides hierarchically structured data in form of a text file. Among other things it is used for data exchange via internet.

Summary

The document introduces the possibilities for data exchange between the EIC and Endure ALPS. Exchanging data between Endure ALPS and the EIC may follow one of two principles. The first principle is a complete data transfer. That might cause performance problems growing with the amount of data. The second principle is to transfer choice data.
Therefore the definition of a shared meta-data model is required. The meta-data model is needed to subscript the data wanted. Several meta-data are already provided by the Endure ALPS. Sharing most of the meta-data does not cause any difficulties, but sharing Endure ALPS' measures is not resolved yet.
To implement a protocol based on meta-data related data transfer several established open standards already exist. The most interesting standards for the needs in the interaction between the ENDURE-ALPS and the ENDURE IC are the OAI-PMH and SRU / SRW. In the current stage of development a complete definition of the protocol using one of the standards is not part of the document. The choice will be made in the group and according to this the protocols will be developed.

1. Fundamentals

The document determines the possibilities for data exchange between the Endure ALPS and the Endure Information Center. To clarify the roles of the participating systems the following definition seems useful. The intention is to provide a chosen set of documents, managed by Endure ALPS, for the public in general. Because Endure ALPS will only provide its content, it acts as a data provider. In contrast to that the Endure Information Center only expects the content delivered by Endure ALPS for further representation. Therefore the Endure Information Center always acts as recipient. Generally any data exchange is performed to one of the following two principles.

1.1. Complete Transfer

The first principle describes a transfer of the complete content. Therefore all datasets are transferred from time to time. Hereby it is also possible to transfer only those documents changed or added after the previous transfer. In such case the content exits twice, on the data providers' side and on the recipients' side. Such a procedure of the data transfer generally comprises a large quantity of data, depending on exchange intervals.
Endure ALPS is not able to comprehend the changes of its content i.e. the system does not keep any kind of history of the previously transferred version and would exclude already transferred datasets. Therefore using such principle always requires a complete data transfer causing larger performance difficulties and therefore it is not recommended.

1.2. Request-oriented Transfer

An alternative is to transfer content individually which means the recipient only requests those parts of the content relevant for any individual query. In this procedure the content is not saved on the recipients' side. The quantity of data transferred directly depends on the EIC users' query. Requesting content directly also ensures the reflection of content changes immediately. The detailed procedure is described in 3. Protocols
For such procedure a common base for data identification is needed. In other words data have to be found to describe the data required by the user. That kind of data is called meta-data. To work with meta-data, the data provider and the recipient has to share the meta-data.

2. Meta-Data

Internally the Endure ALPS (as well as most of the other currently in the development included applications) already uses meta-data for subscribing the content it provides. Most of that meta-data is already adopted as core of a common Endure data scheme. In detail the meta-data used to subscript Endure ALPS' reports contains crops, pests, measures and regions.
Except of regions all of that meta-data are relevant for the EIC. The following items describe the use of the Endure ALPS meta-data for the EIC.

2.1. Crops

All crop entries contained within the common Endure data scheme are based one the EPPO Plant Protection Thesaurus[1]. The EPPO specification identifies almost all cultivated plants and several abstract groups[2] using a unique identifier as well as captions for several languages. The EPPO specification also provides taxonomy for all contained plants.

[1] http://eppt.eppo.org/
[2] Abstract groups means collections like 'waste water pipes' or 'arable land'

Because the EPPO code is well-defined it is useful as part of the meta-data.

2.2. Pest

Another subscripting criterion is the pest entity. As well as crops pests are provided by the EPPO specification. Hence the pests are suited as well as crops to be part of the meta-data.

2.3. Measure

Generally any report describes the use of a certain measure. Therefore the measures are the most important criteria for subscripting. Because measures are dealt with JKI-internal, they are not standardized. Working with measures requires that all measures are known for both sides of data transfer. As well as crops and pests it is recommended to use unique identifiers to identify certain measures. The common Endure data scheme already defines such unique identifiers for all measures. To use the reports by the EIC it has to implement a table containing the measure identifier at least. Besides the table may also contain the caption of the measure if it is required for representing.
Implementing the measures on both sides requires keeping the content of both tables equal. Changes on one table (must concern Endure ALPS' table only) must be communicated to the other side.

2.4. Content

The content is the reports managed by Endure ALPS. Any report describes non-chemical plant protection measures for certain pest-crop-combinations. Therefore any report always contains a summary. Additionally any report also contains attachments comprising the document summarized.

2.4.1. Summaries

Summaries only consist of a title and a text. The text may contain HTML tags for formatting. Because EIC represents the titles and texts within an HTML page, the contained HTML tags should not cause problems. Additionally summaries may also contain bibliographical references. The references only provide information about authors, titles etc. The information only consists of texts which will not cause any problems.
Attachments are not as simple to handle as their summaries. They consist of the document described by the concerned summary. The documents may either be present as uniform resource locator (URL)[3] or as entire file. Whereas the transfer of an URL should not cause any difficulties, the entire files must be examined carefully. Generally Endure ALPS only provides files consisting of grey papers. Those are papers not published yet, which does not mean that they are not intended for public in general. On that score offering a report also requires offering its grey papers if present.
Transferring the entire file to EIC is not useful, because the EIC is not able to process the content of the file. A better solution is to provide an URL for the file locating the internal file from the Endure ALPS system. That mechanism is almost the same as the one providing public attachments. Therefore it does not differ for the EIC whether the file is located within the Endure ALPS system or anywhere else. Thereby difficulties might appear because of security reasons. The Endure ALPS system runs within a protected environment. Accessing its data without authentication must not be possible. The question whether it is sufficient just authenticating EIC as authorised user within this context must be resolved.

[3] mainly as http-address

3. Protocols

To realize a data transfer between different systems, a definition of a protocol is required. The protocol not only defines messages-exchange supplied by the protocol but also the used middleware. Because both of the systems are pure web-applications, the usage of HTTP as transport protocol is useful. In that context HTTP only presents the transport layer, defining the physical submitting of any data. The data itself must be the messages containing either meta-data for requests or summaries and document for transferring.

3.1. Accomplishment

A suitable transfer protocol has not only to allow the transfer of the content; rather it must be capable to submit only the content requested by an EIC user. As already mentioned, the content is subscripted by meta-data. Therefore a practical solution must allow providing the meta-data for the EIC. The meta-data serve as parameter for a query. Endure ALPS does submit its content based on the query of EIC

3.2. Open Standards

Implementing a proprietary protocol is certainly an option but not recommended. Especially for document exchange there are several XML-based generic open standards using HTTP as transport layer. Each of these standards already defines a frame for message-exchange and therefore only a minimum of own definitions and agreements is required. Providing interfaces supporting well-defined established standards also eases the integration of further recipients later.
The person responsible for the EIC (Hugo Besemer, WUR) and the one responsible for Endure ALPS (Alexander Herr, JKI) already discussed several established open generic standards. The most promisingly standards seem to be the OAI-PMH and the SRU / SRW.

3.2.1. OAI-PMH

The OAI Protocol for Metadata Harvesting defines an XML-based protocol allowing the harvesting and processing of meta-data. Main intension of the protocol is to locate publications of different vendors on heterogeneous repositories. Therefore any repository (data provider) provides meta-data describing the publications they offer. All meta-data provided by a repository might be harvested for processing.
From our point of view the OAI-PMH seems useful because of the principle of harvesting and processing of meta-data.

3.2.2. SRU / SRW

Search / Retrieve via URL is a technical standard for bibliographical information systems. Just like OAI-PMH the SRU uses meta-data to describe publications. Because of its descent the meta-data generally contains bibliographical criteria (author, title and so on). Additionally the SRU also provides a query language (CQL) allowing explicitly querying for a certain kind of information. In most aspects the SRU and OAI-PMH offers almost the same features. Concerning the retrieval the systems strongly differ. Whereas the SRU offers a much more granular approach (using complex CGL queries), the main intension of OAI-PMH is to retrieve the data of the provider entirely.
SRW is an extension allowing the data exchange in form of a SOAP-based web service.

4. Conclusion

The integration of Endure ALPS' reports in the EIC is basically possible. The most efficient procedure is to transfer choice data for any EIC user query. Therefore most of the meta-data

are already defined, but the sharing of Endure ALPS' measures may cause problems. Because measures are dealt with JKI-internal, they are not standardized. Therefore it is recommended to use unique identifiers to identify certain measures. The common Endure data scheme already defines such unique identifiers for all measures. The EIC has to implement a table containing the measure identifier at the minimum to use the reports.
To define the data exchange protocol itself, there are already several established open generic standards. Both standards introduced bases on XML as well as the EIC; an additionally Endure ALPS support can be easily implemented.