



atira**information**

Populating PURE with data

Created by: Atira A/S

Date: August 15, 2008

Version: 1.0

Contents

1. Populating PURE with data	3
1.1. Content types in PURE	3
1.1.1. Three ways of populating content types with data	3
1.2. How content types are usually populated	4
1.3. About relations between content types	6
1.4. Importing old publications into PURE	6
1.4.1. Normal process	7
1.4.1.1. Post-import editing	8

1. Populating PURE with data

This document only describes different ways of getting data into PURE. For more information about the system itself, please refer to the other available document.

1.1. Content types in PURE

First a brief recap of what types of content are often in PURE:

1. Persons
2. Organisations
3. Projects
4. Publications
5. Activities
6. Journals
7. Publishers
8. Equipment
9. Funding bodies
10. Funding programmes
11. Student thesis
12. Press clippings
13. External publications
14. Clinical trials

What data is in PURE depends on what data model is used. PURE comes with a default data model that can be used. However, it is completely replaceable and very often universities take the opportunity to define their own data model from scratch or to require modifications to PURE's default model.

No university use all the types mentioned above, but types 1-5 are considered primary in most countries, and the are almost always present.

1.1.1. Three ways of populating content types with data

Populating these types with data in PURE can basically be done in three different ways:

- A) Manual data entry
- B) Dynamic integration
- C) Import

Different tools in PURE and services by Atira are provided for each.

Manual data entry needs no explanation, but PURE offers many tools for making this task effective and fruitful for researchers and other staff members. Please see other available documents about PURE.

We use the following differentiation between dynamic integration and import:

- Import¹ is moving data from other systems, which will go out of production, and into PURE.
- Dynamic integration is setting up either one-way or two-way dynamic communication between a live data source and PURE.

Import in relation with a PURE implementation project is a service delivered in a process that uses PURE's PXA-format and import framework.

Dynamic integration in relation with a PURE implementation project is a system customization resulting in one or more source-specific integration plug-ins² in PURE's plug-in integration architecture.

A PURE project can contain several import projects and several dynamic integrations.

It is obvious that when planning how PURE should be set up at a university, two things should be avoided to the maximum extent possible: Manual data entry and data redundancy.

Well-planned and controlled redundancy can be quite OK, and in lack of better measures like a mature SOA environment, it can also be necessary. In these situations, however, any piece of data must be mastered by one system only. Mastering data from two systems simultaneously is usually difficult to set up effectively.

1.2. How content types are usually populated

Very often, the content types mentioned above are populated with data in PURE as described below:

Persons and Organisations are dynamically retrieved from a live data source such as an LDAP or an Active Directory into PURE.

Projects are dynamically retrieved from a live data source such as a financial system. However, the financial system often contains only little information such as the project name, number, and basic funding info. Therefore, users enrich the project info in PURE; adding info such as relations to Funding bodies and Funding programmes, adding relations to Organisations (where the project is situated at the university), adding relations to Persons (who runs it and takes part in it), adding relations to Publications (what Publications was produced under the project), adding richer descriptions of the project, etc.

1 We call it import and not migration although that might be splitting hairs - in our terminology migration is a form of data processing that might become necessary (almost always is necessary) if data doesn't map to the new data model by default or if data from multiple, inhomogeneous sources must be imported into one.

2 These plug-ins are able to handle one-way or two-way synchronization of data between one source and PURE. They can also be used to enforce business rules; depending on requirements and on the data source in question. See other documents about PURE for more information about this and related areas.

Publications - *Old* publications are normally subject to data migration from one or more previous system that are to go out of production and into PURE. This is done one time in history, and after that *new* publications are created in PURE by researchers and other staff members using PURE's different tools for that (tools for efficiency, data quality, rights management, etc. See other available documentation).

Journals and Publishers - We have two cases now, one German and one Danish, where the two types are going to be imported into PURE from official government resources containing "approved" journals. Approved means that if a publication is published in one of those, it will increase the government funding for that university. Technically, these government resources are quite different: The German one is a rating list in the form of an Excel spreadsheet, the Danish one is a database exhibiting itself over the OAI-protocol in a standard format. In the first case the sheet is simply imported into PURE annually by a small, very cheap custom import feature created for the purpose in PURE, and the latter is automatically retrieved by PURE's OAI harvesting client when new content is available.

Of course, Journals and Publishers can be created manually in PURE, too. Functionality exists in PURE for supporting manual creation of journals simultaneous with the publication registration process while still controlling journal data quality strictly (avoiding doublets, etc.).

Activities are usually entered manually into PURE by researchers and their staff. Since it is a relatively new content type, it usually does not exist anywhere else in the data environment at a university. The need for researchers to enter it manually goes well in hand with the fact they usually are motivated to register such activities and appreciate the opportunity. Originally, the content type was added to PURE's default data model by request from researchers themselves.

Equipment is not the most used type. Research institutions that already runs a register likes to integrate that register to PURE dynamically in order for users of PURE to be able to create relations to other content types in PURE and to run reports.

Funding bodies and Funding programmes can be created in PURE by either of the three methods. It is not a content we see very often, and defining a norm is not possible. However the tendency seems to be that they come from central national or even international sources, from which dynamic integration or at least annual imports are possible.

Student thesis - some universities find they belong naturally in their PURE system, some take the opposite position (that is why it is an optional module). In the first case, theses from previous years are usually imported from one or more old systems into PURE. After that, the former system or systems are taken out of production, and new student theses are created directly in PURE - but not by means of PURE's user interface: Most universities do not want to grant students access to PURE. Instead, students submit theses onto a web page or an intranet page, from which PURE then imports the data periodically. Also, the PUREportal framework includes the option for setting up an access-controlled page for that purpose. Further, PURE's web service can be used to deliver data about submitted theses to a student admin system. One Danish university is currently very interested in this.

Press clippings are usually imported regularly from a press clipping service. Such services usually deliver some kind of structured data feed, typically some kind of XML format.

External publications are always entered manually or retrieved by the authors themselves using PURE's Self-import module (provided the desired publication is available at some source supported by the Self-import module).

Clinical trials is a content type only used by the pharmaceutical industry and hospitals. Usually, protocol numbers are retrieved dynamically by integration to the local CTMS-systems and then enriched in several ways through a complex approving workflow before published and stored according to FDA and EudraCT requirements.

The above information is also illustrated here:

No.	Description	Manual data entry	Dynamic integration	Import
CONTENT TYPE				
1	Persons		X	
2	Organisations		X	
3	Projects		X	
4	Publications (old and new, respectively)	X		X
5	Activities	X		
6	Journals		X	
7	Publishers		X	
8	Equipment	X		
9	Funding bodies		X	
10	Funding programmes		X	
11	Student thesis	X		
12	Press clippings		X	
13	External publications	X		X
14	Clinical trials		X	

Table 1: How different content types are usually populated in PURE

1.3. About relations between content types

Very briefly: Crucial to the whole purpose of setting up a PURE solution are the relations between the content types. These relations make reporting, searching, specific exports, and browsing possible, and they solve a number of data model challenges to; name changes to authors and organisations, to mention one. Please also see other available documents.

1.4. Importing old publications into PURE

Very often the biggest import project when implementing PURE is importing old Publications. Old publication data is often on one or more older systems in a data structure not identical to that in PURE.

One particularly important task when importing old publication data is creating correct relations between publications and other content types such as Organisations and Persons (the author relation).

1.4.1. Normal process

First, Organisations are put into PURE. Normally they come from a live source, and in that case a dynamic integration is set up to that source using a source-specific synchronization plug-in on PURE's integration plug-in architecture. PURE is initialized with organisations the first time the dynamic integration runs.³

Usually, Persons come from the same source as Organisations (e.g. an LDAP or an Active Directory), and in that case the Person-Organisation relations are quite easy to create. It is perfectly possible to integrate Organisations and Persons from two different sources, but in that case valid identification of Persons' relations to Organisations must exist in the data. Otherwise it is not possible to have dynamic integration of these two important content types from two different live sources.

Once Organisations and Persons are in PURE with correct relations, the old Publications can be imported. Sufficient information to establish relations between Publication and Organisations must be present in the import data.

If needed, we can use a script with an algorithm to translate organisation names in the publication data to variations that can be used to find the correct matches in PURE's organisation names.

If this does not work for all organisation names in the publication data, we will simply send an Excel spreadsheet back to the university with the names we could not match and ask what that organisation is called in the original source (e.g. the LDAP or the Active Directory that PURE organisations originates from).

After that, each publication in the import data is properly matched to the Organisations in PURE.

Next, we proceed to creating relations between Publications and the Persons in PURE; the Author relation.

If there are sufficient identification of Persons in the import data, these relations are created effortlessly already at import. It must be identification such as a valid personnel-ID that will match the same or a mappable ID in PURE's person information (which is the same as the LDAP's or the AD's person information).

If no such valid identification is in the import data, we temporarily import all authors as "External authors" (a state normally reserved for co-authors not employed at the university).

Our service can stop here. It would be perfectly possible for the university to have the researchers, departmental secretaries or other employee groups log into PURE and change the state for "External authors" to "Internal Authors" by finding and choosing the proper persons in PURE.

However, we often complete this work for the university by using a script with an algorithm to create possible variations of the author name in the import data and find all possible matches in PURE. Then, these possible matches are exported to an Excel spreadsheet that goes to the university. This Excel file contains columns with Publication title, Author name from the import data, and one Person from PURE that possibly possible matches. There is one row per possible match - so if the publication's author name was A. Millington, and if there were 7 Millingtons starting with an "A" among the PURE Persons, there will be seven rows in the Excel file. The

³ The normal principle for these integrations is one-way synchronization of data from the source to PURE, thereby keeping an always-update copy of data in PURE by running the synchronization daily, hourly, or whatever fits the requirements. However, if the live source is set in a modern, able SOA-environment, PURE will be fully able to integrate in ways that will avoid the redundancy.

last column is reserved for the university to mark with "Yes" or "No", thereby identifying the Person in PURE that is identical to the author name for that publication in the import data.⁴

Once we get the filled-out Excel file back, we correct the state of authors in PURE based on that. Then, the import is complete.

1.4.1.1. Post-import editing

Once the import is complete, one question remains: Would the university like the authors or other members of staff to be able to log into PURE and modify the publications?

Publications can be in several states. This is customizable, but most universities uses three states and three roles: Entered, Approved and Validated. Researchers enter, Organisational Publication Editors approve (like an institutional secretary), and Central Publication Validators validate (they are often librarians at the research library).

If there are no wish to edit publications after import, they are imported in the state "Validated". If librarians should be able to modify them, they can just as easily be imported in the state "Approved". If also institutional or department secretaries should be able to modify them, they can be imported in the state "For approval". And if researchers should be able to modify them, the state "Entered" would be right.

Anyway, publications can always be "re-validated" in the original workflow by the top authority in the workflow allowing it.

⁴ Please note that since we have Organisations in place in PURE, the Excel file will be sub-divided by Organisation, allowing the university to split that task of filling out the file among the proper Schools, Faculties, Institutes, Departments and so on. This eases the work considerably.