EUROPEAN HOLOCAUST RESEARCH INFRASTRUCTURE

European Holocaust Research Infrastructure
H2020-INFRAIA-2014-2015
GA no. 654164

Deliverable 10.2

Collection Description Publishing Services

Henk van den Berg
DANS-KNAW

René van Horik
DANS-KNAW

Start: May 2015 [M1]
Due: April 2017 [M24]
Actual: April 2017 [M24]

EHRI is funded by the European Union

Document Information

Project URL	www.ehri-project.eu
Document URL	[www……]
Deliverable	D10.2 Collection Description Publishing Services for Archives
Work Package	WP10
Lead Beneficiary	P1 DANS-KNAW
Relevant Milestones	MS2
Dissemination level	Public
Contact Person	René van Horik / rene.van.horik@dans.knaw.nl / +31623297389
Abstract (for dissemination)	This deliverable reports on the “Metadata Publishing Tool” (MPT) that is aimed to provide a sustainable solution for the publishing and maintenance of metadata provided by a CHI to the EHRI portal. The text in the deliverable consists of a selection of the documentation of the MPT software that is available at the site that contains the complete documentation of the Metadata Publishing Tool, at: <http://rspub-gui.readthedocs.io/en/latest/index.html>
Management Summary	N.a.

Table of Contents

1 Introduction

Deliverable 10.2 is described in the DoA as follows “*The collection description publishing services will provide the facilities for archives to publish information about their Holocaust relevant collections on their institutional websites according to standard protocols and the needs of EHRI. The publishing services will include a user-interface to manage the publishing of the collections. This deliverable will also design and implement the facilities for acquisition and ingest of the published archival descriptions for use in the EHRI infrastructure.*”

The work done in this task can be summarized as follows: The service developed in this task is named Metadata Publishing Tool (MPT). The basic function of the service is to facilitate the synchronisation of local metadata with metadata stored in the EHRI portal. The service automatically processes new records, modified records and deleted records in the local metadata information system. The tool implements the ResourceSync framework Specification[1].

Detailed documentation (61 pages) of the MPT service is available online at: http://rspub-gui.readthedocs.io/en/latest/index.html.

The MPT service is a graphical user-interface and application that is deployed on your laptop or local workstation. It comes in the form of a wizard that helps a content-savvy, but technically unskilled archivist to import, select and filter EAD-files, group them in sets, create ResourceSync metadata files over changes, additions and deletions in each set of EAD-files, publish the sets of EAD-files and the ResourceSync metadata on the web site of the CHI and finally to assess and verify the URL’s on the institution web site.
The source code of the publishing service can be found in two gitHub repositories: https://github.com/EHRI/rspub-core and https://github.com/EHRI/rspub-gui .
Executables for various operating systems can be downloaded from the releases page of the rspub-gui github repository: https://github.com/EHRI/rspub-gui/releases

The ‘facilities for acquisition and ingest’ - in short the Destination software - can be found at the github repository rs-aggregator: https://github.com/EHRI/rs-aggregator. The deployment at the EHRI portal of the Destination software requires alignment with the “services integration” task of WP10 and is in progress.

[1] The ResourceSync specification describes a synchronization framework for the web consisting of various capabilities that allow third-party systems to remain synchronized with a server's evolving resources. The capabilities may be combined in a modular manner to meet local or community requirements. This specification also describes how a server should advertise the synchronization capabilities it supports and how third-party systems may discover this information. The specification repurposes the document formats defined by the Sitemap protocol and introduces extensions for them.See: <http://www.openarchives.org/rs/toc> [Cited March 26, 2017]

2 The MPT Software and the EHRI data infrastructure

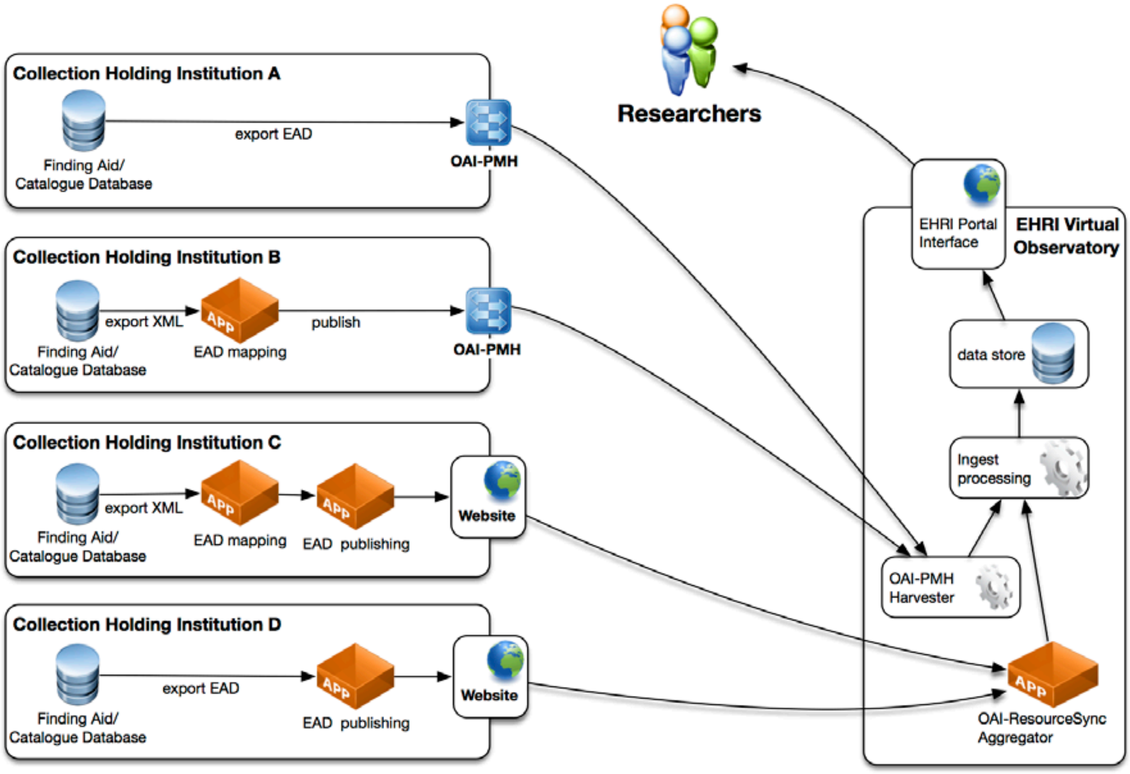


Figure 1. EHRI data infrastructure

The EHRI project has developed software services to assist the data integration process. The software services are represented by the orange boxes in Figure 1.

To what extend the service is usable for a CHI depends on the way the local data infrastructure is organised. E.g whether metadata on archival holdings are available in a digital form, its format and how the information infrastructure is able to communicate with the outside world.

Two standards form the core of the data integration workflow: (1) the EAD metadata format[2] for archival finding aids and (2) the OAI-PMH protocol for metadata harvesting[3].

Within EHRI we make a distinction between several types of information infrastructures.

- CHI type A = CHI that can export metadata in the EAD format and supports the OAI-PMH metadata harvesting protocol, so the EHRI harvester can automatically gather

[2] EAD stands for Encoded Archival Description, and is a non-proprietary de facto standard for the encoding of finding aids for use in a networked (online) environment. Finding aids are inventories, indexes, or guides that are created by archival and manuscript repositories to provide information about specific collections. While the finding aids may vary somewhat in style, their common purpose is to provide detailed description of the content and intellectual organization of collections of archival materials. EAD allows the standardization of collection information in finding aids within and across repositories.
The EAD Metadata Schema and information related to the EAD standard can be found at: <https://www.loc.gov/ead/> [cited 24 March 2017].

[3] See: <https://www.openarchives.org/pmh/> [cited 14 July 2017]

- the metadata from the CHI.
- CHI type B = This CHI supports the OAI-PMH harvesting protocol. The metadata itself, however, is not available in the EAD format. A local format is used that can be exported in XML. A tool is available to convert the local metadata format into EAD. This is the EAD Conversion Tool
- CHI type C = This CHI does have metadata available in a local format. So the metadata has to be converted to the EAD standard. For this a tool is available. The CHI does not have a OAI-PMH data provider installed. EHRI has developed a metadata publisher (the Metadata Publishing Tool (MPT-tool), covered in Deliverable D10.2) that implements the ResourceSync Framework. This framework describes a synchronization framework for the web that allows third-party systems to remain synchronized with a server's evolving resources.
- CHI type D: is capable of exporting metadata in EAD format, but does not have a OAI-PMH service. So it also needs the MPT-tool (that supports the ResourceSync framework.

The Metadata Publishing Tool (MPT) service is part of the EHRI data infrastructure as illustrated in Figure 1. The MPT service implements the “EAD Publishing” and “OAI-ResourceSync Aggregator” functions as used by CHI type C and CHI type D

3 Downloading the MPT software

The MPT Software can be downloaded from Github at: https://github.com/EHRI/rspub-gui/releases/tag/1.0.rc.6

3.1 System requirements

The MPT software runs under Windows and Mac OS operating systems. Details on installment can be found at: http://rspub-gui.readthedocs.io/en/latest/rst/rsgui.install.html

4 About

Metadata Publishing Tool (MPT) is a desktop application that facilitates the publishing of resources and sitemaps in conformance with the ResourceSync Framework Specification.

Metadata Publishing Tool, rspub-gui and rspub-core were developed by Data Archiving and Networked Services (DANS-KNAW) under auspices of the European Holocaust Research Infrastructure (EHRI).

The documentation in this rtd is intended for end users and system administrators.

- Download latest executables from the releases page of the rspub-gui project.
- Source location: https://github.com/EHRI/rspub-gui.
- The GUI is based on rspub-core. See https://github.com/EHRI/rspub-core.
- In case of questions contact the EHRI team.

4.1 ResourceSync

The ResourceSync Framework Specification describes a synchronization framework for the web consisting of various capabilities that allow third-party systems to remain synchronized with a server's evolving resources. More precisely the ResourceSync Framework describes the communication between Source and Destination aimed at synchronizing one or more resources. Communication utilizes http and an extension on the sitemap protocol, an xml-based format for expressing metadata, relevant for synchronization.

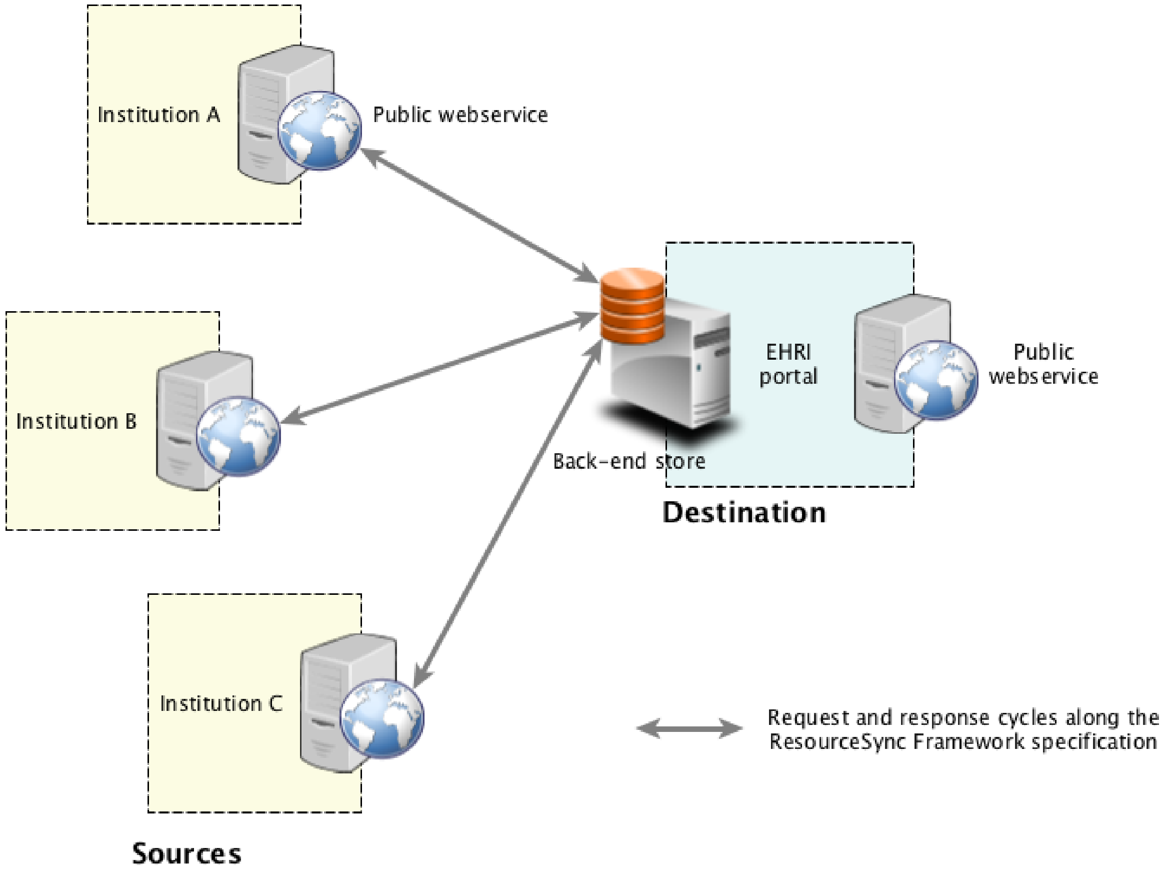


Fig. 2. External logistics. The ResourceSync Framework Specification at work.

Collection Holding Institutions expose content and ResourceSync metadata on their web servers. The central hub (in this case the EHRI Portal) is actively collecting resources and keeping them in sync with the aid of published sitemaps.

We can say that the ResourceSync Specification is a perfect fit for solving the external logisticswhen it comes to synchronizing resources between a central Destination and various Sources. Figure 2. depicts the external logistics.

When the resources we are trying to synchronize are not web-resources by them selves but instead stem from information systems, databases or other places within an organization, we are faced with other problems, which we can qualify as related to internal logistics.

4.2 Metadata Publishing Tool

Metadata Publishing Tool is an application that solves various problems related to the internal logistics:

- How do we collect and import resources from various places within the organization;
- How do we select relevant resources;
- How do we create ResourceSync sitemap metadata on relevant resources;
- How do we export resources and sitemaps to the web server;
- How do we verify that the exposed URL's are correct and our ResourceSync site ready to be harvested by a Destination.

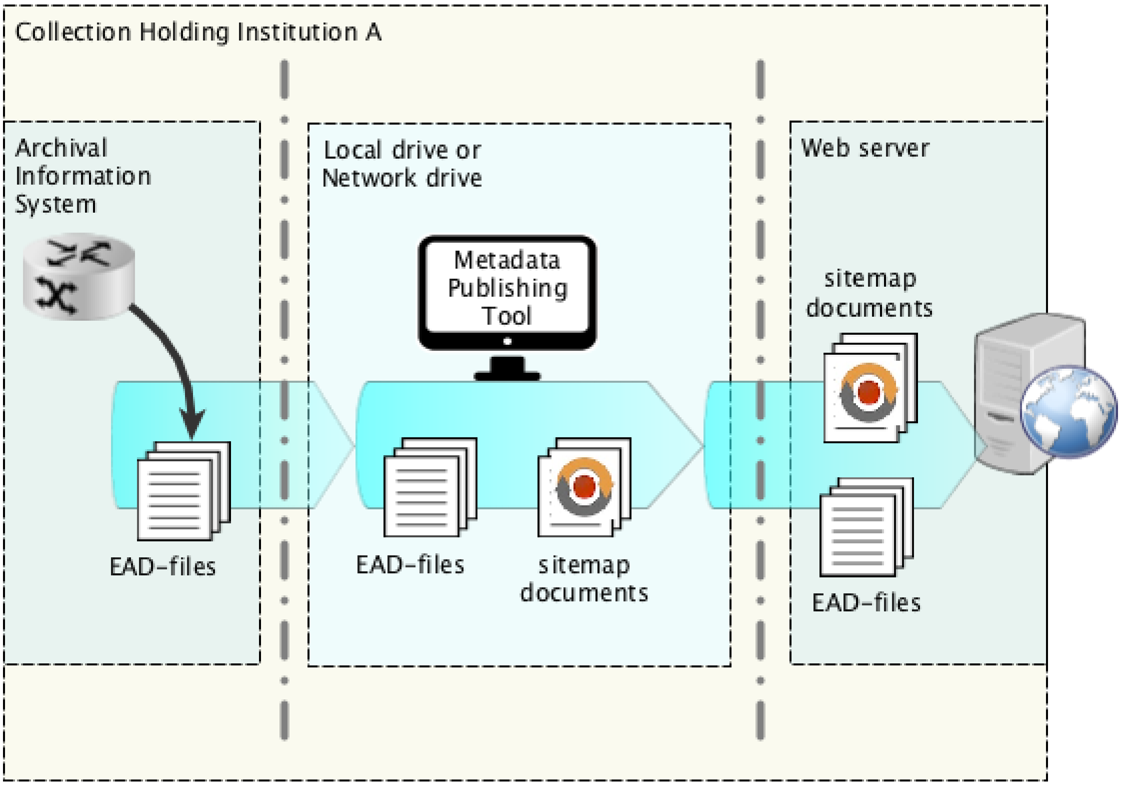


Fig. 3. Internal logistics. Metadata Publishing Tool at work.

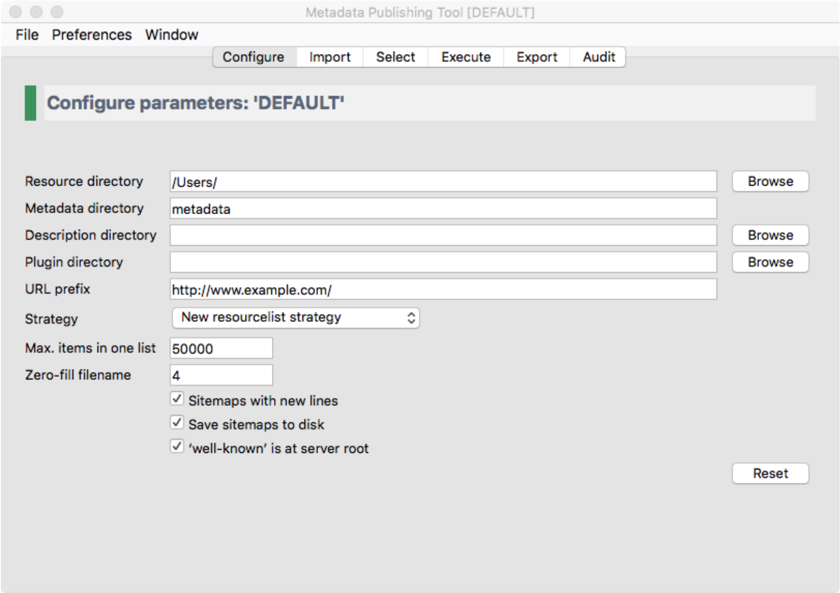
Figure 3. depicts internal logistics and the role of Metadata Publishing Tool. The situation described may be exemplary for Collection Holding Institutions (CHI's) within the EHRI infrastructure, although different situations may equally be applicable. Metadata Publishing Tool is an application that is deployed on your laptop or local work station. From there you collect and select resources, create the ResourceSync sitemaps, export resources and sitemaps to your web server and verify the exposed URL's.

Configuration of Metadata Publishing Tool may need the hand and insight of a technically skilled person. Once configured it can be managed by archivists and other content-savvy users that do not necessarily have technical skills.

4.3 Interface of the MPT service

Below screenshots of the graphical user interface of the MPT service are provided. They are taken from the extensive documentation of the service, available online at: http://rspub-gui.readthedocs.io/en/latest/

These screenshots give an indication of the functionality of the service.

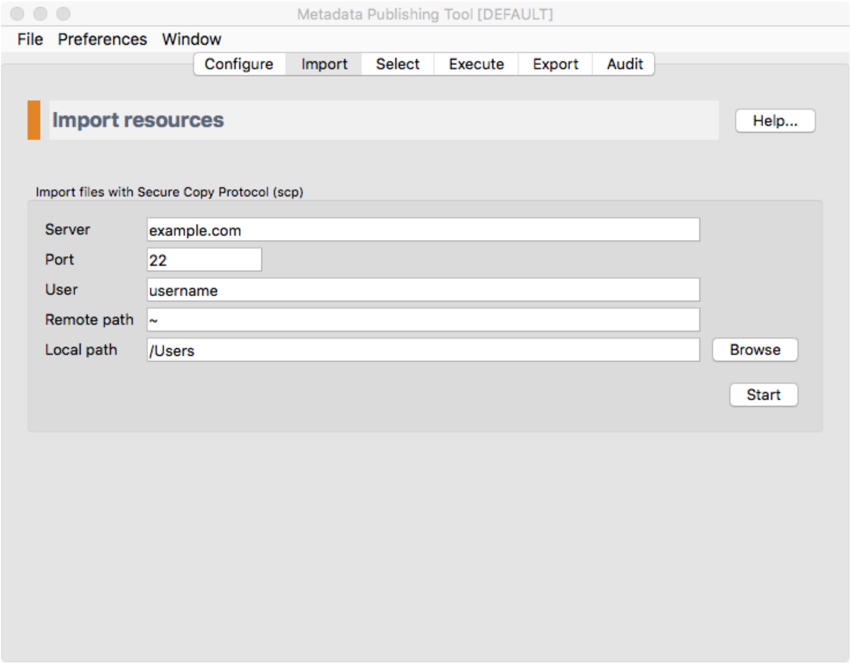


Configure parameters: 'DEFAULT'

Resource directory	/Users/
Metadata directory	metadata
Description directory	
Plugin directory	
URL prefix	http://www.example.com/
Strategy	New resourcelist strategy
Max. items in one list	50000
Zero-fill filename	4

☑ Sitemaps with new lines
☑ Save sitemaps to disk
☑ 'well-known' is at server root

Screenshot of the Configuration page

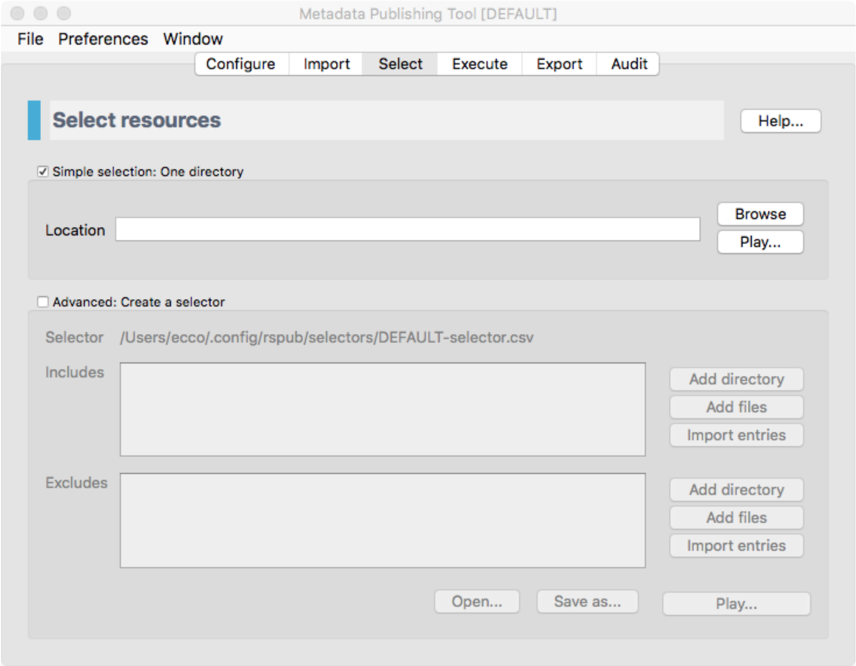


Import resources

Import files with Secure Copy Protocol (scp)

Server	example.com
Port	22
User	username
Remote path	~
Local path	/Users

Screenshot of the (mport resources page



Select resources

☑ Simple selection: One directory

Location

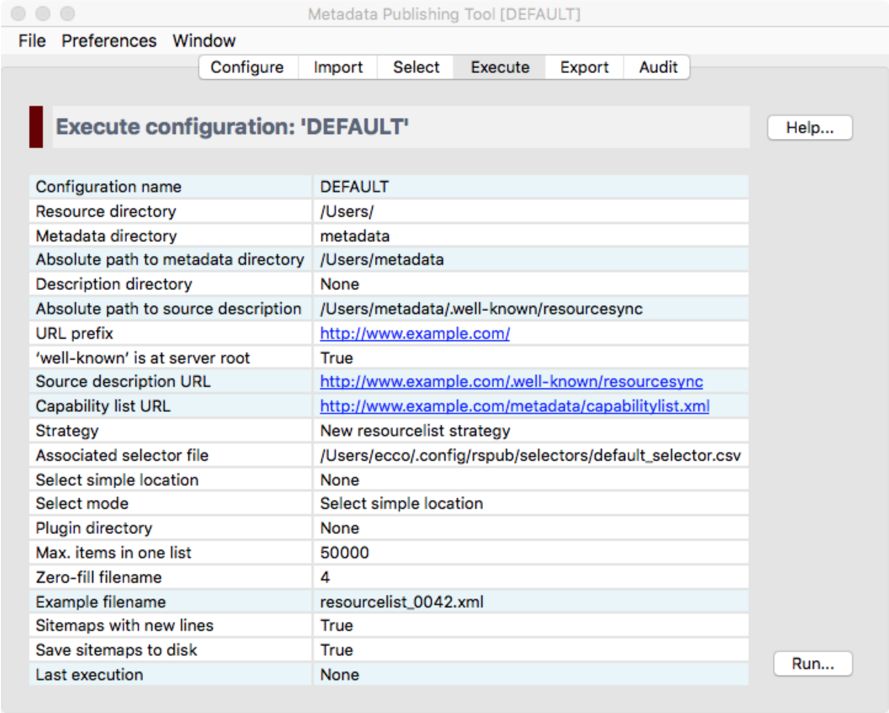
□ Advanced: Create a selector

Selector /Users/ecco/.config/rspub/selectors/DEFAULT-selector.csv

Includes

Excludes

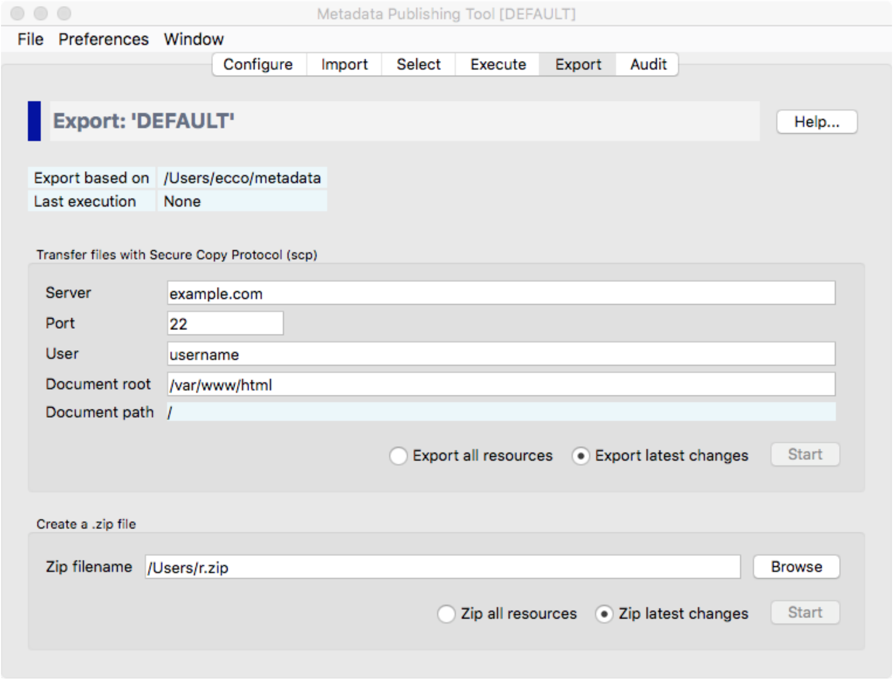
Screenshot of the Select page


Metadata Publishing Tool [DEFAULT]
File Preferences Window
Configure Import Select Execute Export Audit
Help...
Run...

Execute configuration: 'DEFAULT'

Configuration name	DEFAULT
Resource directory	/Users/
Metadata directory	metadata
Absolute path to metadata directory	/Users/metadata
Description directory	None
Absolute path to source description	/Users/metadata/.well-known/resourcesync
URL prefix	http://www.example.com/
'well-known' is at server root	True
Source description URL	http://www.example.com/.well-known/resourcesync
Capability list URL	http://www.example.com/metadata/capabilitylist.xml
Strategy	New resourcelist strategy
Associated selector file	/Users/ecco/.config/rspub/selectors/default_selector.csv
Select simple location	None
Select mode	Select simple location
Plugin directory	None
Max. items in one list	50000
Zero-fill filename	4
Example filename	resourcelist_0042.xml
Sitemaps with new lines	True
Save sitemaps to disk	True
Last execution	None

Screenshot of the Execute page



Metadata Publishing Tool [DEFAULT]
File Preferences Window
Configure Import Select Execute Export Audit
Help...
Browse
Start
Start

Export: 'DEFAULT'

Export based on	/Users/ecco/metadata
Last execution	None

Transfer files with Secure Copy Protocol (scp)

Server: example.com
Port: 22
User: username
Document root: /var/www/html
Document path: /

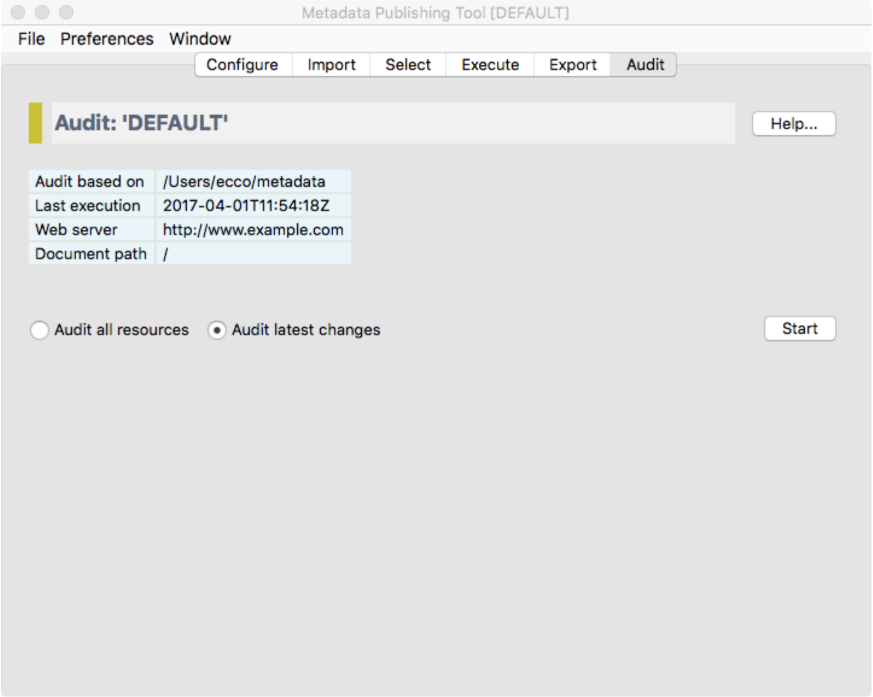
○ Export all resources ● Export latest changes

Create a .zip file

Zip filename: /Users/r.zip

○ Zip all resources ● Zip latest changes

Screenshot of the Export page



Metadata Publishing Tool [DEFAULT]
File Preferences Window
Configure Import Select Execute Export Audit
Help...
Start

Audit: 'DEFAULT'

Audit based on	/Users/ecco/metadata
Last execution	2017-04-01T11:54:18Z
Web server	http://www.example.com
Document path	/

○ Audit all resources ● Audit latest changes

Screenshot of the Audit page