CESSDA Workplan: Metadata Harvesting Tool

Ørnulf Risnes (NSD)
John Shepherdson (UKDS)

EDDI 15, Copenhagen
3 December 2015

UK Data Service

CESSDA

Consortium of European Social Science Data Archives

- Bring together social science data archives across Europe
  http://cessda.net/National-Data-Services/CESSDA-Members

- Developing a pan-European Research Infrastructure (RI)
  - Facilitate researcher access to important resources of relevance to the European social science research agenda regardless of the location of either researcher or data

Commissioned by CESSDA

- 2015 work plan launched RI development
- Metadata Harvester task is part of the work plan

  "The objective of this Task is to select and/or develop and implement into CESSDA service a metadata harvesting tool that will enable the efficient compilation and operation of the CESSDA Product and Service Catalogue, the CESSDA Secure Access Portal, and other data management tasks and data supply services."

- Open Source bundle due Q2 2016

Task Objectives

- Produce an easy to use metadata harvesting service
- Extensible design - use plugin architecture for inputs and outputs
  - wide range of metadata sources must be harvested
  - data to be emitted in a variety of metadata standards (more to life than DDI!)

Delivery Partners

- NSD and UKDS
  - Design, implement, quality assure and document the metadata harvester to produce an Open Source bundle
  - NSD lead the Task
- FSD, SND, DDA
  - Test the OS project - build additional input/output plugins and harvest a variety of metadata source types and languages

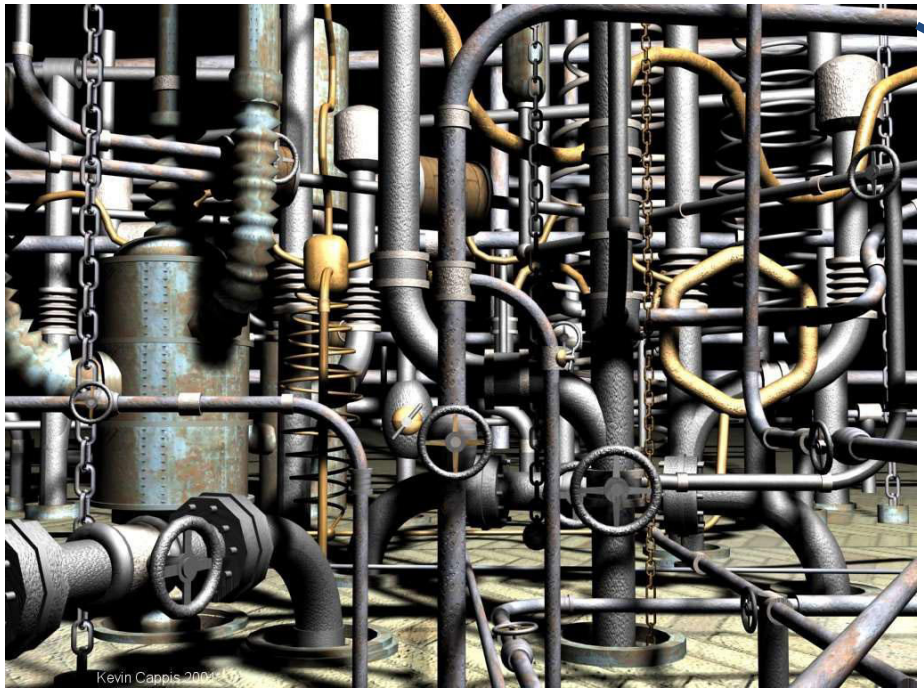
Foundations

- Build on outputs of Data without Boundaries WP12
  - Prototype Resource Discovery Portal (DwB-RDP)
  - Harvests DDI 1.2 from CESSDA SP's Nesstar servers and converts to DwB-Disco format
  - See harvesting ingest system report and DwB Resource Discovery Portal description for more details

- Needs to interoperate with CESSDA metadata model
  - yet to be defined

Harvester as complex system

Metadata sources

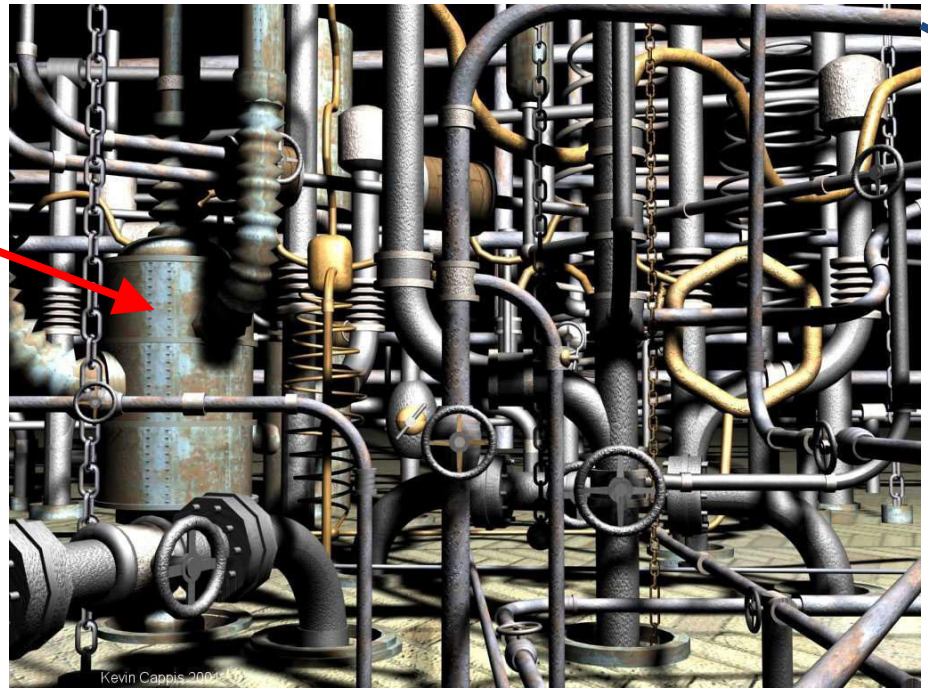


Consumers

Harvester as complex stateful system

Data store

Metadata sources



Consumers

Images sourced via Google Images

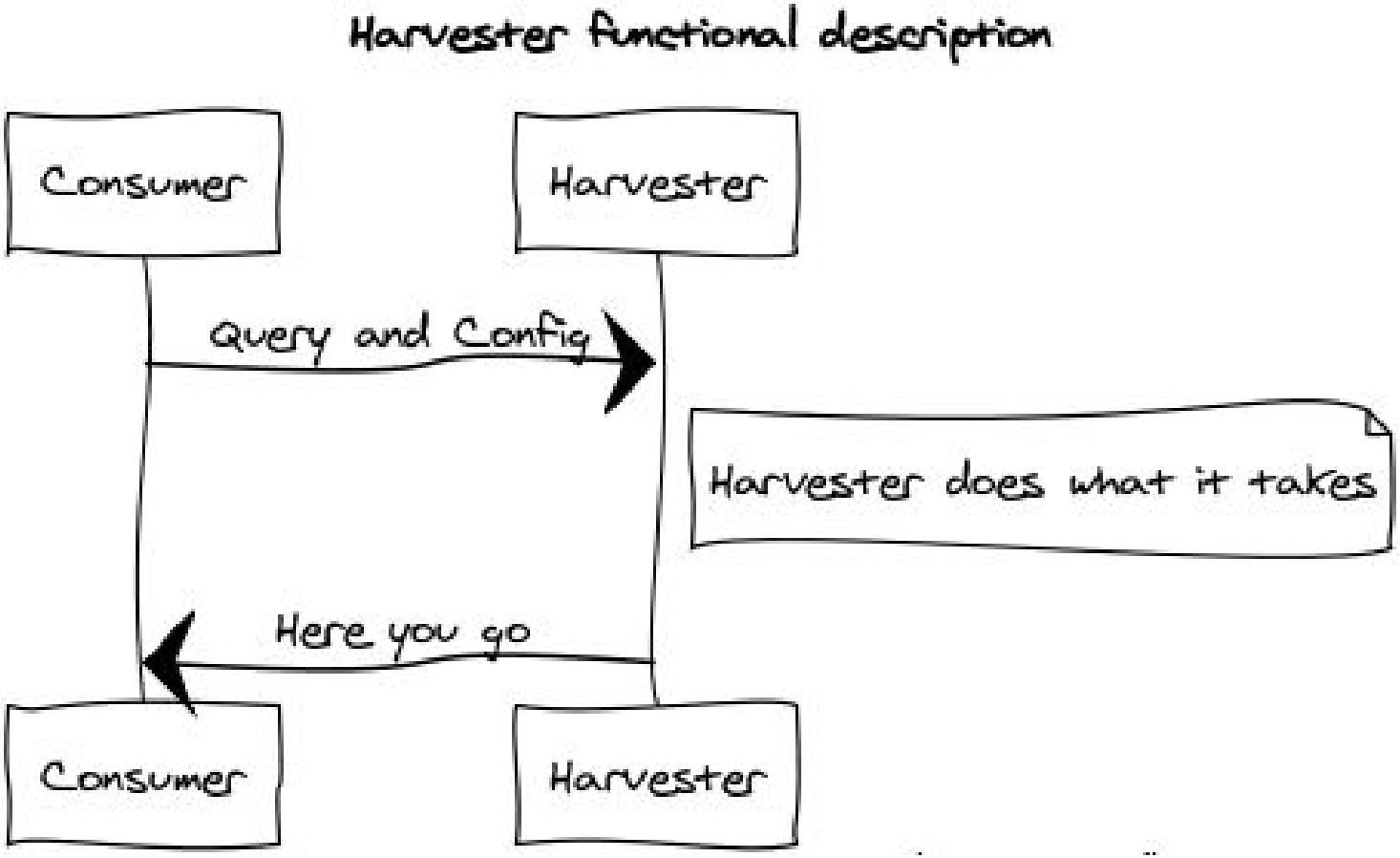
Metadata Model

- Based on RDF-Disco
- How is it different?
  - will perform gap analysis

- Harvesting challenges and normalisation
  - Simplifying assumption
    - normalisation handled by Consumer
    - resumption and completeness are responsibility of Consumer

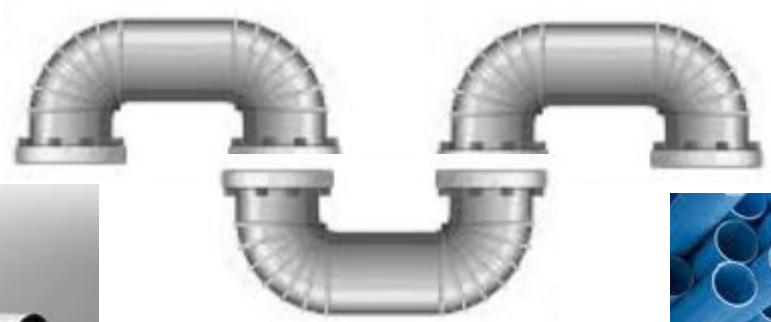
=> Harvester as stateless system

Functional description



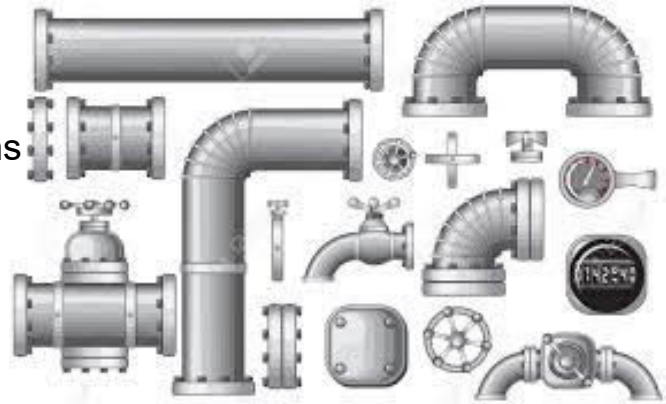
Harvester as simple adaptor

Metadata sources



Consumers

Harvester plugins



Images sourced via Google Images

UK Data Service

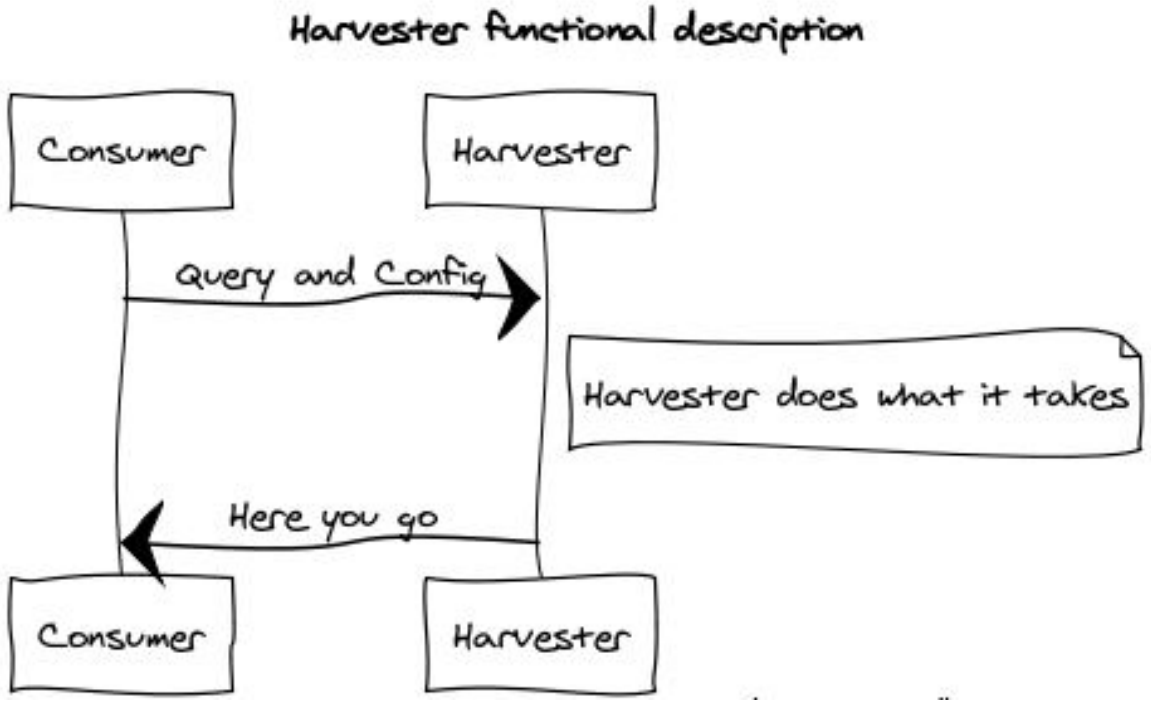
Harvester Extension Mechanism

- Webhooks

  - Harvester calls a provided URI
  - URI handles the call

Images sourced via Google Images from pusher.com

Functional description



Queries:

- ListRecordsForRepository
- GetRecordFromRepository

Arguments:

- Repository/Record URI
- Repository type
- Type handler URI (optional)

Success Criteria

- Impact Analysis
  - How will it affect
    - CESSDA Service Providers,
    - CESSDA ERIC,
    - EU Researchers

- Quality/Usability of OS bundle
  - Establish maturity rating using NASA Reuse Readiness Levels, prior to testing
  - Testing undertaken by FSD, SND, DDA against System Usability Scale

NASA Reuse Readiness criteria

Ten levels for each of following:

- Documentation
- Extensibility
- Intellectual Property
- Modularity
- Packaging
- Portability
- Standards Compliance
- Support
- Verification and Testing

- *Security*
- *Internationalisation and Localization*

Deliverables

- Metadata harvester as a service
  - Provides an API for clients to consume it
- Administration tool
  - Used to monitor and manage the harvester service
  - May be readable to many, but will be writable by few
- Publically available Open Source Bundle
  - Code base and documentation
  - Facilitates creation of new harvesters and output formatters

Questions

How will it be developed?

Where will it run?

Development Environment

- Common, cloud-based tool chain
  - lower barriers to entry for all Service Providers
  - no need to install and configure locally
    - code repositories
    - automated build and test
    - documentation area

- Enure CESSDA has access to
  - source code
  - configuration files
  - technical documentation

that underpin its products and services

Production Environment

- Short-term cloud based hosting
- Experience will feed in to CESSDA's requirements for compute and storage in order to host and run the components of the Research Infrastructure

Questions

Thanks for your attention