Capturing Experiments in the Wild

Mark Hedges, Shrija Rajbhandari, Stella Fabiane

King's College London

Open Repositories 2010, Madrid, 6th July 2010

www.kcl.ac.uk

A message for our funders

Funded by JISC (Joint Information Systems Committee) in UK

JISC

Part of the JISC Information Environment programme

Runs for 2 years: 1st April 2009 – 31st March 2011

Overview

Aim: to develop a research data repository for certain research groups

Context: the scientists, their data, their research

Objectives, issues and constraints

Provenance

Work so far

Context

Who are the researchers?

Randall Division of Cell and Molecular Biophysics

Cross-disciplinary research unit in School of Health Sciences

Research groups:

- Macromolecular crystallography (e.g. proteins)
- Nanoimaging (microscopes, 3D, time lapse)
- NMR (Nuclear Magnetic Resonance)
- Molecular Simulations and Bioinformatics
- others …

What do they do?

Raw data captured from equipment

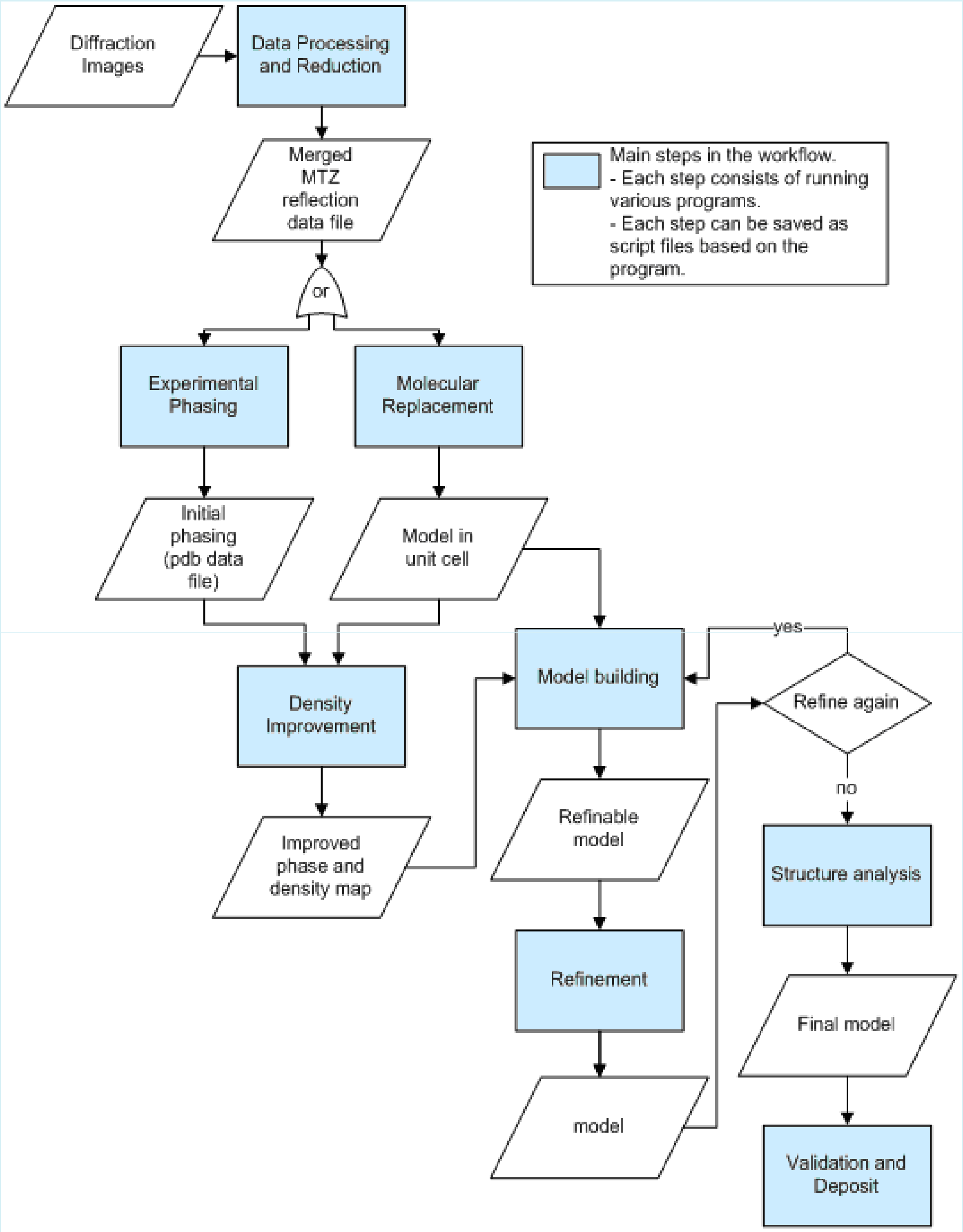
Transferred to desktop machine

Processed/analysed on desktop

At the end:

- Published output (articles, final data objects)
- Rest of data either archived to DVD in drawer or discarded.

Different specific processes, common patterns



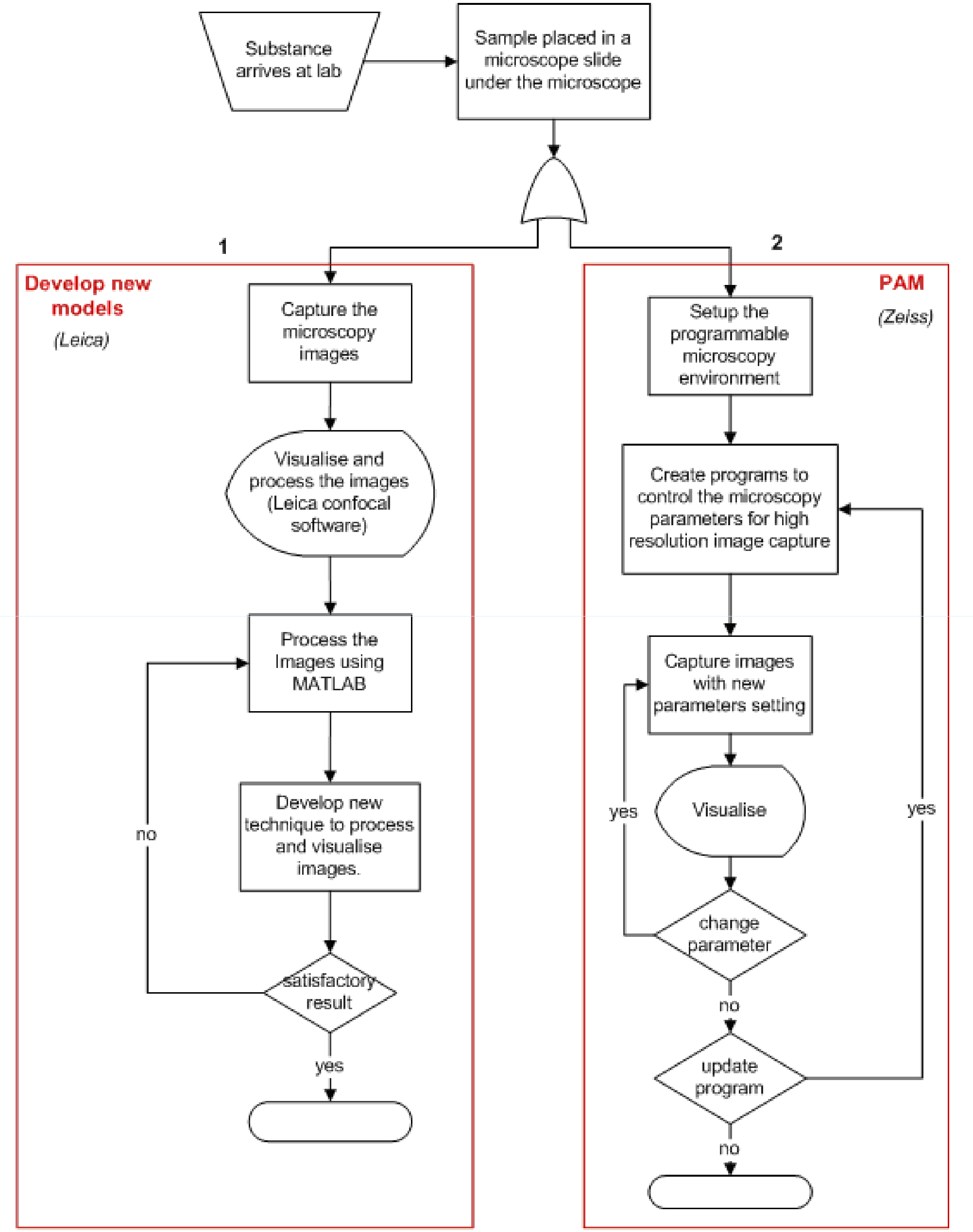
🟦 Main steps in the workflow.
- Each step consists of running various programs.
- Each step can be saved as script files based on the program.

Example – protein crystallography workflow

Input: set of diffraction images from a protein crystal

Output: protein structure object published to PDB (and articles)

Many intermediate stages and data files, which tell the full story of the experiment


Substance arrives at lab
Sample placed in a microscope slide under the microscope
1
2
Develop new models
(Leica)
Capture the microscopy images
Visualise and process the images (Leica confocal software)
Process the Images using MATLAB
Develop new technique to process and visualise images.
satisfactory result
no
yes
PAM
(Zeiss)
Setup the programmable microscopy environment
Create programs to control the microscopy parameters for high resolution image capture
Capture images with new parameters setting
Visualise
change parameter
yes
no
update program
yes
no

Example: Nanoimaging

- Input: microscopy images (e.g. from cell sample, including 3D, time lapse, …)
- Output: processed images, algorithms (e.g. software, MATLAB scripts, …)
- Many intermediate stages and data files, which tell the full story of the experiment

Objectives

What do they/we want to do?

Basic requirement – repository for data

Integrate with researchers' processes (tools, equipment) – avoid "deposit"

Capture experimental structure (provenance)

Trace processing to validate published results

[Model for use with other researchers]

[Interface to institutional preservation infrastructure]

Issues / Constraints

Not an entirely integrated environment

Specialised tools (mostly) running on desktops

Tools (mostly) interact only with the file system

Some proprietary; some open source, developed by scientists and shared

Researchers' workflows not very predictable

An experiment can generate lots of files (1000+)

As little as possible change to how researchers work (avoid additional burden)

On the bright side

A given researcher works through an experiment at the same desktop machine (simplifies our initial work)

Much information about files and process can be captured from (e.g.) file headers, log files

Much work already done in these disciplines on metadata, vocabularies

Prototype implementations

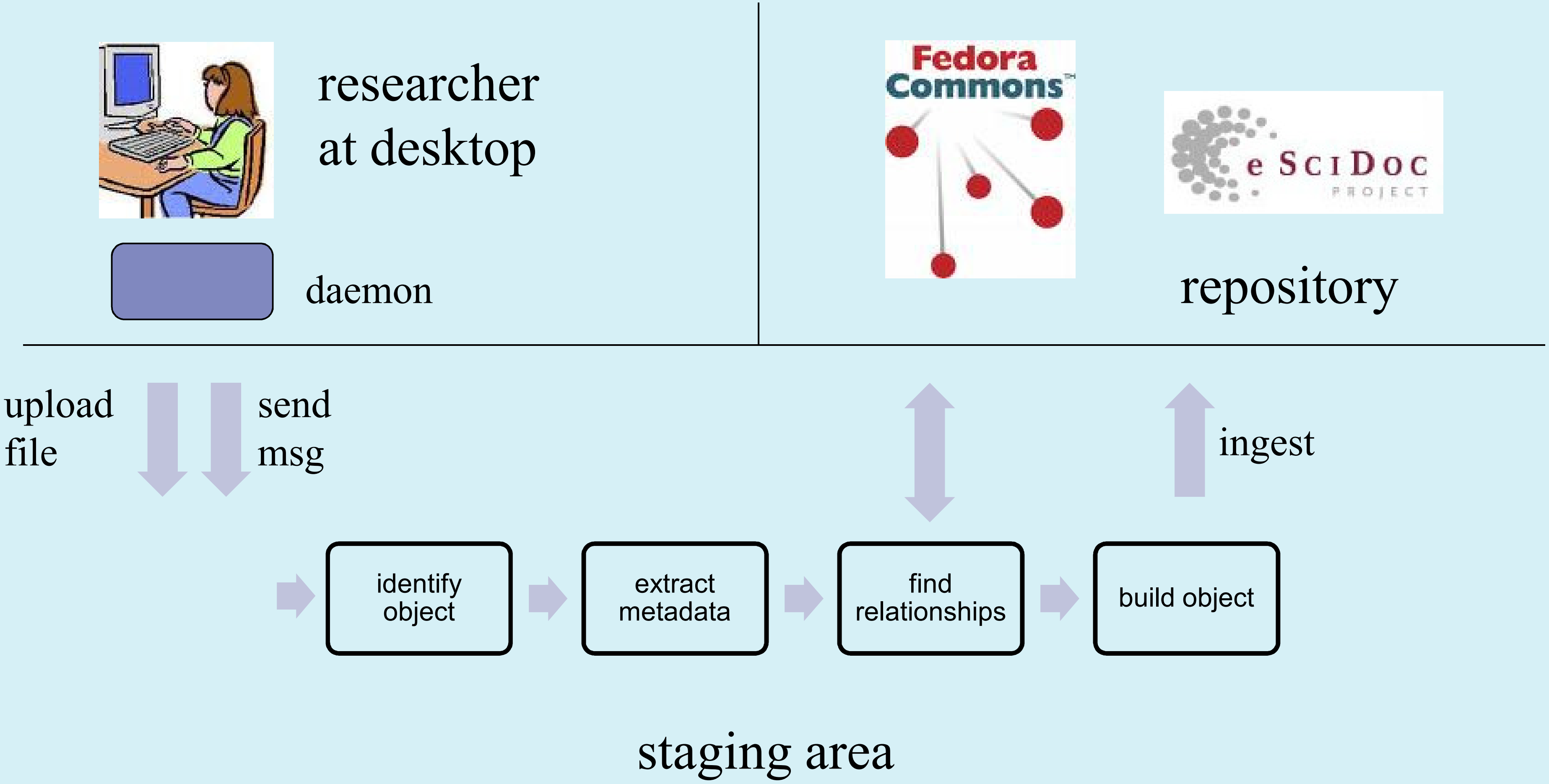
First prototypes – focus on data capture and organisation

Limit scope to two research groups

Approach taken:

- “scavenge” at desktop – automatic capture of as much data/metadata as possible
- ingest workflow that interprets captured information and creates objects for repository
- researcher supplements this information

Data capture



researcher at desktop

daemon

Fedora Commons

eSciDoc PROJECT

repository

upload file

send msg

ingest

identify object

extract metadata

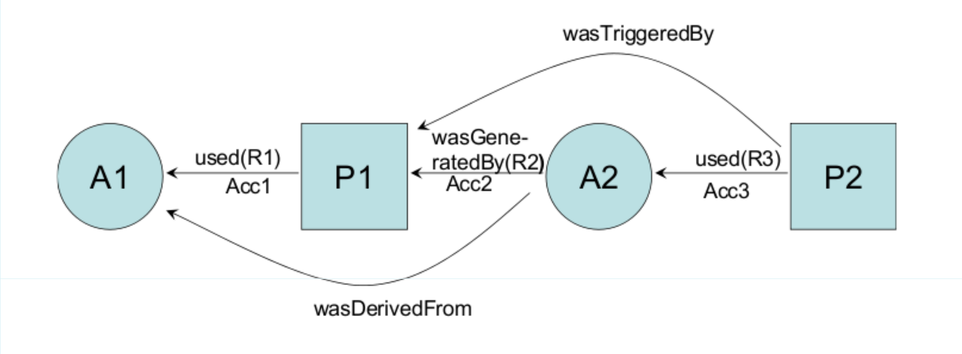
find relationships

build object

staging area

Provenance

Open Provenance Model



3 node types: artifact, process, agent

5 edge types: used, generated, triggered, derived, controlled

Can be extended by adding annotations to nodes/arcs

Choice of granularity/focus (e.g., artefact or process-centric)

[This and following 2 slides derived from OPM presentations]

Nodes

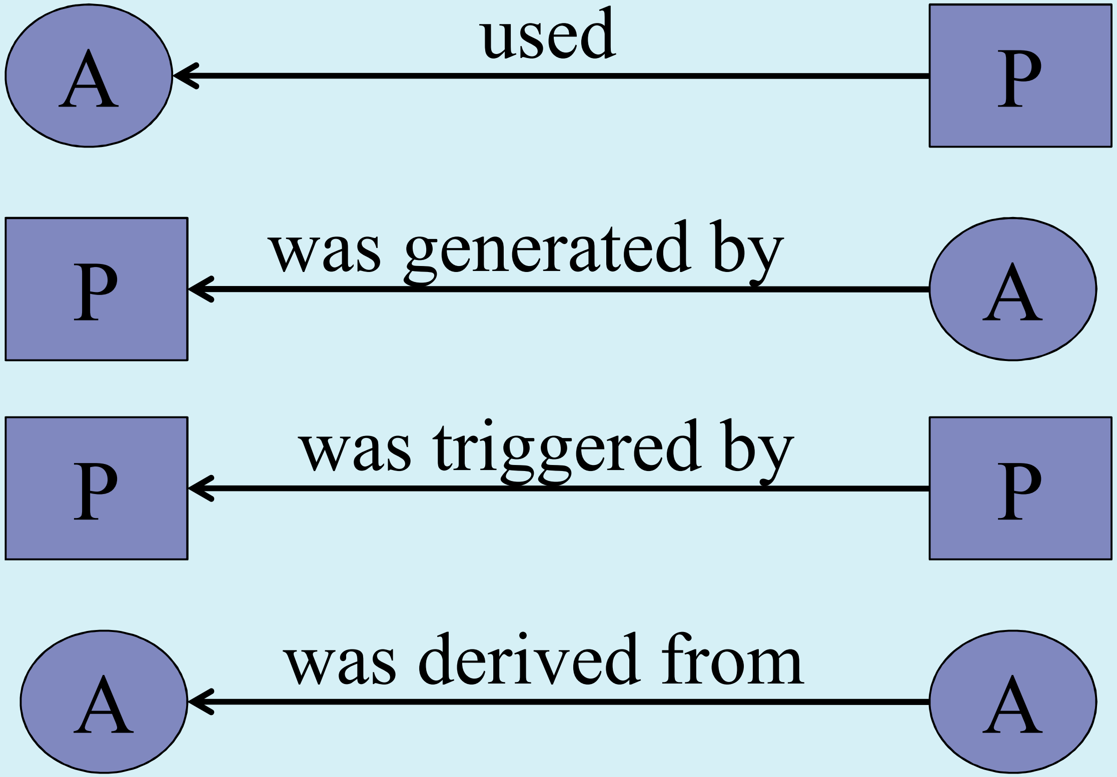
Artifact: Immutable piece of state, which may have a physical embodiment in an physical object, or a digital representation in a computer system.

Process: Action or series of actions performed on or caused by artifacts, and resulting in new artifacts.

Agent: Contextual entity acting as a catalyst of a process, enabling, facilitating, controlling, affecting its execution.

Edges

Specify how nodes are related



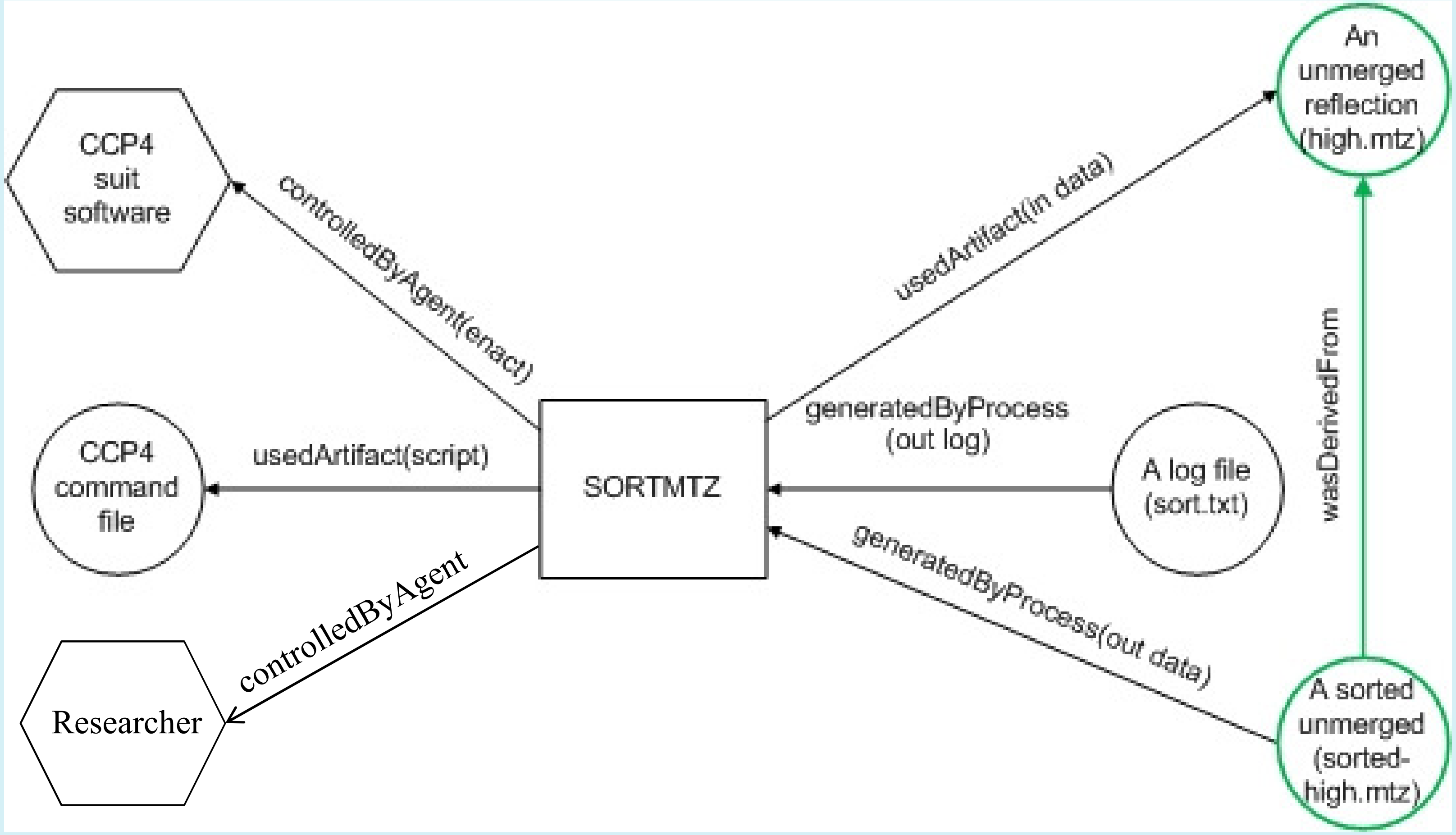
In our case

Artifact: captured file

Agent: researcher, software/tool used

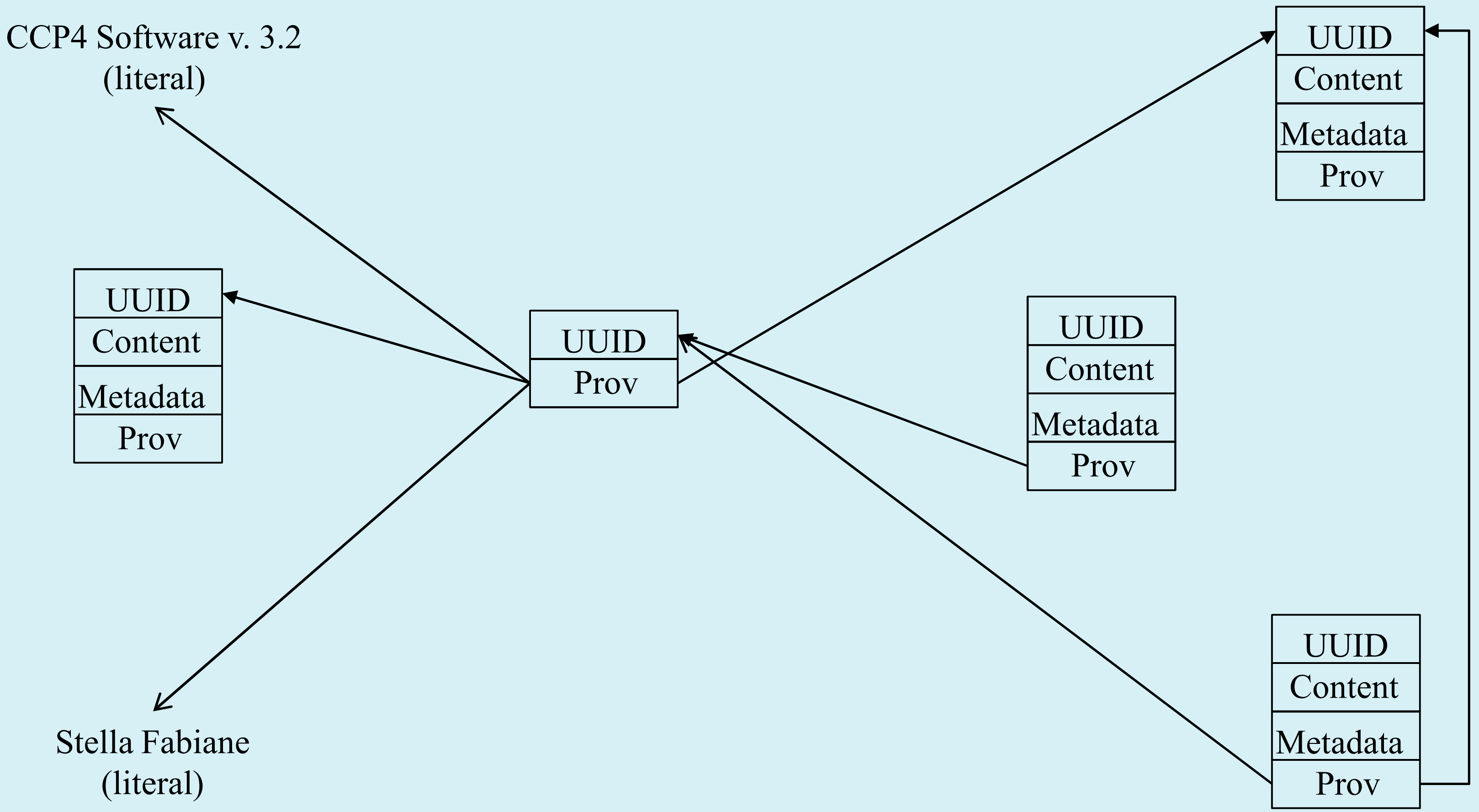
Process: researcher uses software to process some files, resulting in some derived files or other new files

Example



CCP4 suit software
controlledByAgent(enact)
CCP4 command file
usedArtifact(script)
SORTMTZ
controlledByAgent
Researcher
usedArtifact(in data)
An unmerged reflection (high.mtz)
generatedByProcess (out log)
A log file (sort.txt)
wasDerivedFrom
generatedByProcess(out data)
A sorted unmerged (sorted-high.mtz)

Modelling in repository



So far ...

Where are we?

Speaking to researchers, finding out what they do, what they want, what they won’t accept

Describing these research processes in a form that we can use

Analysis of datasets and metadata

Small-scale prototyping

Experiences

Figuring out what a researcher does can be painful and time-consuming

Lack of comprehension on the part of (some) researchers

Complexity of interpreting files and their relationships (messy evidence)

What next?

Working prototypes for two research groups

User interface

Publication – articles, data

Use of OAI-ORE concepts

Extension to other research groups

Contacts

mark.hedges@kcl.ac.uk

shrija.rajbhandari@kcl.ac.uk

stella.fabiane@kcl.ac.uk

http://bril.cerch.kcl.ac.uk/