SeGoFlow:
A Semantic Governance Workflow Tool

Sven Lieber✉, Anastasia Dimou, and Ruben Verborgh

IDLab, Department of Electronics and Information Systems, Ghent University – imec
{firstname.lastname}@ugent.be

Abstract. Data management increasingly demands transparency with respect to data processing. Various stakeholders need information tailored to their needs, e.g. data management plans (DMP) for funding agencies or privacy policies for the public. DMPs and privacy policies are just two examples of documents describing aspects of data processing. Dedicated tools to create both already exist. However, creating each of them manually or semi-automatically remains a repetitive and cognitively challenging task. We propose a data-driven approach that semantically represents the data processing itself as workflows and serves as a base for different kinds of result-sets, generated with SPARQL, i.e. DMPs. Our approach is threefold: (i) users with domain knowledge semantically represent workflow components; (ii) other users can reuse these components to describe their data processing via semantically enhanced workflows; and, based on the semantic workflows, (iii) result-sets are automatically generated on-demand with SPARQL queries. This paper demonstrates our tool that implements the proposed approach, based on a use-case of a researcher who needs to provide a DMP to a funding agency to approve a proposed research project.

Keywords: Provenance, Workflow, Governance, Data Management

1 Introduction

Funding agencies and other institutions require data management plans (DMPs) before research proposals can be approved, see e.g. H2020[1], FWO[2], NWO[3]. These DMPs are questionnaires asking questions like *What data will you collect or create?*, or *How will you ensure that stored data are secure?*. The aim of these plans is to assess if the performed data processing complies to FAIR[4] principles and force the creator to reflect over the planned data usage within a project. A DMP is only one example of a document describing data processing in a project. Another example are privacy policies. When the study includes the collection and

[1] H2020 Programme Guidelines on FAIR Data Management in Horizon 2020, Europe.
[2] Fonds Wetenschappelijk Onderzoek – Vlaanderen, Belgium.
[3] Nederlandse Organisatie voor Wetenschappelijk Onderzoek, The Netherlands.
[4] Findable, Accessible, Interoperable, Re-usable.

processing of personal data, the consent of the participants is needed. Therefore, a privacy policy needs to be written and handed out to the participants.

Even though DMPs and privacy policies are different documents, yet both aim to describe certain aspects of data processing. However, information of multiple domains is needed to create a DMP or a privacy policy, For instance the legal domain, *which data are considered as personally identifiable information?* Or the technical domain, *are used data stores accessible by third parties?* Users might request or discover all this information, compile it and add it to the DMP, the privacy policy or to any other document a stakeholder requested. Each change of the data processing entails a repetition of that process and creates new versions of all documents, which introduces even more complexity.

This demo shows our data processing workflow editor. Instead of using dedicated tools to generate DMPs or privacy policies, the proposed tool *SeGoFlow* assists in the creation of data processing workflows. *SeGoFlow* shows that a single semantic description of data processing workflows and result-sets, generated by SPARQL queries over that representation, can be used as base for DMPs.

Since not everyone who models data processing is familiar with Semantic Web technologies, our proposed tool makes use of a graphical workflow language to model the data processing, which also improves the communication between stakeholders [1]. The graphical workflow components, that are described with the OPMW ontology[5], can also be reused, which simplifies the need to request or discover relevant information when modelling data processing of a certain project. The screencast available at `https://www.youtube.com/watch?v=6zTRL1WUL5g`, demonstrates the generation of a DMP with our tool.

2 Demonstration

This section introduces our proposed tool, *SeGoFlow*, based on a scenario.

Scenario A researcher performs a study to investigate stress in peoples daily life. Therefore each participant receives multiple wearable devices used to measure values like heart response or skin conductance. The study also involves the processing of participants private data, like their email address. Before the research starts, a DMP needs to be provided to the funding agency. The researcher uses our proposed tool *SeGoFlow* to describe the data processing tasks.

Projects Our tool allows to create graphical data-driven *workflows*. Multiple workflows can be grouped into a *project* to increase readability of each workflow. The introduced researcher creates one workflow to describe the process of data collection, and another workflow describing the data sharing. Figure 1 shows the data collection workflow of the demonstration.

[5] `http://www.opmw.org/model/OPMW/`.

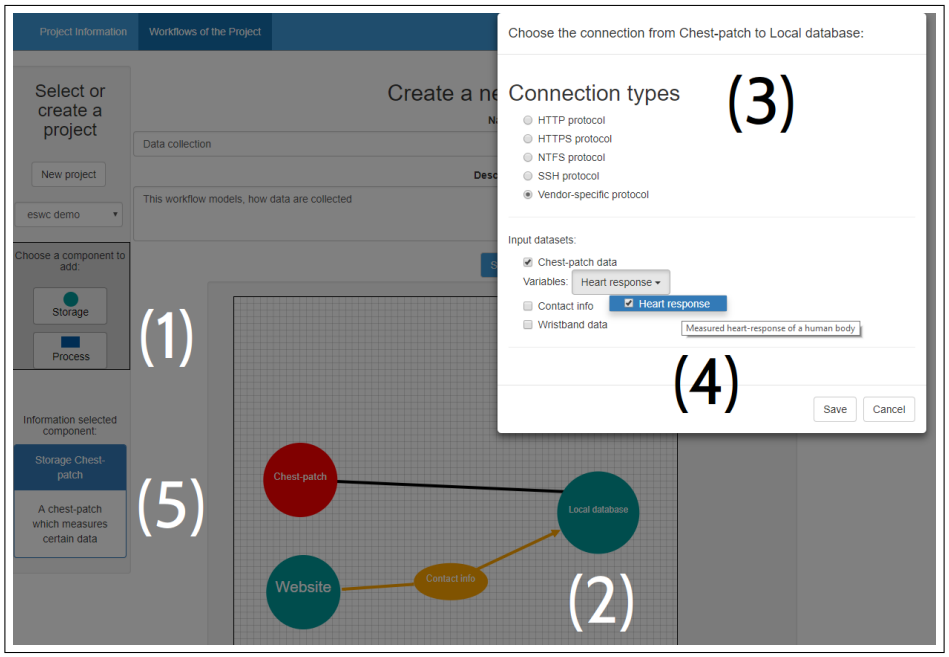

Fig. 1. A screenshot of the user interface. (1) The menu where users can select *data stores* and *processes*. (2) The workflow drawing area in which the workflow is modelled. (3) A connection dialog, asking a user *which* data should be transferred from *chest-patch* to *local database* and *how* it should be transferred. (4) Tooltips when hovering over elements. (5) Detailed information of the currently selected component.

Workflow Components The basic components of a workflow are *data stores* and *data processes*. These can be selected from a *repository* and inserted into a *workflow*, part (1) of figure 1 shows the components menu. Each component is described semantically to express detailed information of multiple domains in an interoperable way. For example data stores, the semantic model describes their physical location as well as persons and organizations having access. The separation of defining a component and re-using a component, allows users with domain knowledge to create and maintain components, whereas users interested in modelling data processing can focus purely on the workflows. Users with the aim of modelling their data processing are provided with more details about workflow components (part (4) and (5) of figure 1).

The introduced researcher finds relevant data stores and processes in the repository and re-uses them. Part (2) of figure 1 shows the drawing area of a workflow, containing the three *data stores Chest-patch*, *Local database* and *Website*.

Connections SeGoFlow offers the functionality to connect selected components of a workflow with each other. One connection symbolizes the flow (transfer) of

data. Since the way *how* data are being transferred carries important semantics, these connections are modelled semantically as well. A connection defines which data are being transferred, and also which protocol or communication channel is used. The tool prompts the user to select data and a transfer protocol. These semantically defined data and protocols are available through a repository as well. Part (3) of figure 1, shows the dialog where a user has to choose the semantics of a connection. In case of the screenshot, the connection from *Chest-patch* to *Local database* is selected. The connection itself is represented as yellow arrow, containing the name of the transferred dataset.

Data management plan generation Based on the workflows of a project, a DMP can automatically be generated via a button of the project description view. Figure 2 shows an excerpt of a generated DMP. The question regarding used data stores could be answered based on the used workflow components (see part (1) of figure 2). Semantics of data transfers (drawn connections) were used to answer the question regarding the secure storage of data (see part (2) of figure 2). Finally the semantically encoded legal domain knowledge that *Email address* is considered as personal data, and the technical domain knowledge that the used data store is accessible by a user of another institution, contributes to the answer of data sharing restrictions (see part (3) of figure 2).

3 Conclusion and Future Work

With *SeGoFlow*, researchers have a semantically enhanced and a graphical representation of data processing. Such a graphical representation allows, for example, researchers to communicate their research to colleagues or other stakeholders. Researchers create detailed workflows which reflect on their data processing. Workflow components, as well as projects and workflows themselves are described semantically in a provenance-aware way that reduces the overhead and provides valuable information in case of e.g. an audit.

A possible future direction is to further assist users in creating workflows. For instance, to provide directly privacy-related feedback when connecting components, as already proposed for computational workflow tools [2,3].

References

1. Johnston, W.M., Hanna, Millar, R.J.: Advances in dataflow programming languages. ACM Computing Surveys (CSUR) **36**(1) (2004) 1–34
2. Gil, Y., Cheung, W.K., Ratnakar, V., Chan, K.k.: Privacy enforcement in data analysis workflows. In: Proceedings of the 2007 PEAS workshop. (2007) 41–48
3. Gil, Y., Fritz, C.: Reasoning about the appropriate use of private data through computational workflows. In: AAAI Spring Symposium: Intelligent Information Privacy Management. (2010)

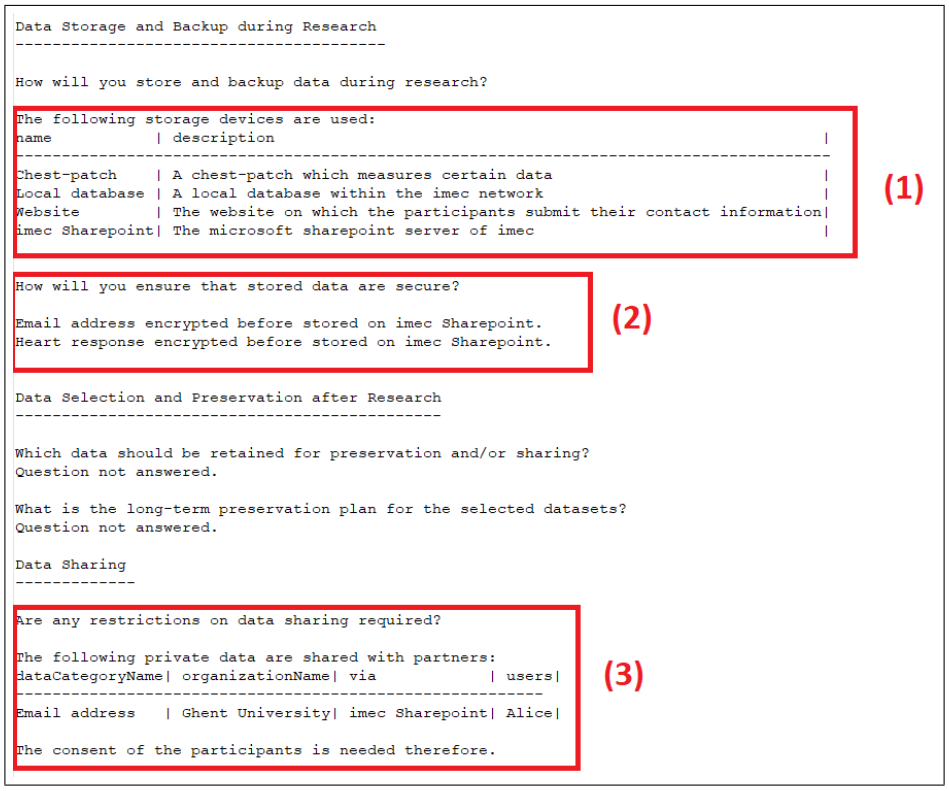

Fig. 2. Excerpt of a data management plan generated by our proposed tool. (1) The answer regarding used data stores is answered based on used workflow components. The semantics of drawn workflow connections contributed to answer questions regarding data sharing (2, 3). The DMP is generated by a text template and SPARQL queries against the workflows created in the demonstration available at `https://www.youtube.com/watch?v=6zTRL1WUL5g`.