2 Intro

First, a little bit of a context. I finalised my transfer report in 2018 and it was reviewed by the present team. Back at that time I was involved with Theme E group 5 and our focal point was upcoming GDPR and issue of data sharing between the parties. Originally we were interested in automated data sharing agreements (or smart contracts) between 2 data controllers. Transparency in GDPR was the main driving factor and the idea was to use the semantic web technologies together with blockchain. Blockchain based smart contracts was hot topic at the time.

After the report, the focus has changed - now rather than focusing on smart contracts, we try to establish a way to produce open, on-demand contract. Having in mind the importance of GDPR defined actors, we started focusing not just on the transparency but the interoperability as well. That ruled out blockchain, also thanks to your feedback in the original transfer report.

The novel idea is to keep using semantic web, utilising its power, but to also utilise BPMN for process modelling. Why BPMN, the reasoning coming in next few slides. Finally as both BPMN and Semantic web ontologies can be serialed through XML, we decided to utilise the XML schema for certain aspects of the model, eventually leading to Schematron as our validation tool.

Final note, back to early last year, we were planning to apply for the continuation hearing, but due to the events it kept being delayed. So rather than waiting, the report was used as a base for the thesis itself and kept growing; soon original 10 pages became 70 pages. I tried to mark the focal points in the thesis so it would be easier to review, but ended up writing chapter abstracts, or summary. I apologize for the possible repetition in the full text. These abstracts, or summaries would not be part of the thesis, we would need further abbreviation, but they will be utilised for the paper that is coming this April, enabling to push this research out for a peer review.

3

This is the structure of this presentation, resembling the structure of the thesis/report itself

4

Main motivation was obviously GDPR regulation, new law that brings a lot of changes how data is shared, processed and stored. Most significant part of GDPR was that it clearly defined actors in the data sharing process, namely data subject, data controller, data processor and also another actor that is not part of the data sharing process itself, but is there, appointed by the state to oversee the process and issues of compliance.

We can not understand GDPR without understanding the roles of the actors. Most significantly, GDPR introduces a number of data subject rights that are counterbalanced by data controller duties. Understanding the actors, their roles, duties, rights and their relationship is a key to understanding GDPR itself. That brings us to the issue of interoperability and the role it plays in our proposed model.

As previously mentioned, transparency is one of the main GDPR requirements, as the data subjects have a right to know all the details about their data, where it is used, for what purpose and who else has access to it.

It is not just the data subject that needs to be aware of this, but, if required, the supervisory authority should have an easy overview of the entire process as well.

That brings us to exploring a possibility of open contract, that can be assessed on-demand at any time and that shows the current state of the process and state of the data.

5
This is the interoperability reference model on the left, showing interoperability points; and then the problem scope on the right - we focus on data exchange between 2 data controllers, primary data controller and the third party that can be another data controller or data processor. This is based on previous work of theme E and the paper produced as a consequence.

6
Second driving factor of the thesis is the open data technologies and how they can be utilised in terms of transparency and interoperability. Semantic web is the first such technology, with obvious advantages of flexibility (which is not direct requirement, but tremendous help in achieving the goal). Associated concept of linked data, especially open linked data helps us with transparency requirement.

To have our model utilised by both legal personnel and the data managers, we look into the business process modeling. One solution seems to be industry adopted and becoming de facto standard in business process modeling - BPMN (Business process model and notation). It also feeds into our transparency requirement, being developed by the OMG group, standardised and completely royalty free.

The commonality between the 2 is what the usually refer to as human readable technologies, in effect meaning what I already mentioned - can be utilised even by a non-technically minded. But effectively the technologies are not limited in that aspect and have the background and potential to be as technical as the needs arise.

Last piece of the puzzle so to speak, is the technicality behind the data exchange, utilising XML or JSON, again both open data technologies, to create 2-way api that captures the current state of the data sharing process

7
That brings us to our research question: To what extent can existing open web data technologies support interoperable GDPR compliance for data sharing between parties?

We break the question down int 4 research objectives:

8
Next part of the thesis is the state of the art, we're looking at semantic web technologies in this slide and will focus on that due to time constraint; in the report we look into other concepts such as data sharing agreements, XML and JSON, right expression languages, etc. ODRL seemed like a most suitable ontology that natively deals with actors, assets, rights and duties/obligations and prohibitions, and it has a concept of profiles- basically extending the core odrl vocabulary.

Linked data platform provides us with restfull interface in data exchange. LDP defines a set of rules for HTTP operations on web resources, using same verbs as we see in http protocol - POST, GET, PUT, PATCH and DELETE.

DUV ontology helps us keep reliable track record following best practices on the web

And finally SHACL tackles the issue of cardinality of the elements and artefacts

DPRL is an umbrella to all these technologies. In essence DPRL is on the most basic level an ODRL profile that expands ODRL to be more specific in terms of the actions definitions that match GDPR terms, clarifying them and removing ambiguity. DPRL is also a framework that is used for defining elements of our open contract.

9
We will quickly jump to another chapter to look into dprl profile details, as due to time constraints we can not discuss it in too many details in this presentation. DPRL utilises Agrements, Assets and Actions classes from ODRL, but also defines more specific Roles class that is a sub-class of ODRL Party class. It also uses SHACL to define constraints and resolve the issue of cardinality that ODRL core is lacking in.

10
The slide shows formal relation between DPRL profile, odrl, dataid and duv

11
Moving back to our requirements chapter, this is a full snapshot of the SySML requirements diagram. it is available for q/a session in full resolution, if needed

12
But to give quick summary of the diagram, on the very top level, as it can be derived from our background and motivation earlier in this presentation; apart from terms frequently used in previous slides - transparency and interoperability, we also have a legal compliance as a top-level requirement

Breaking them down, our model should be presentable, in order to facilitate the target audience as discussed earlier

Should follow strict business logic

Because it is open contract that can be assessed on-demand, it should be possible to validate it dynamically

And of course due to transparency requirement, it should follow open standards

13
This is a continuation of the requirements model, in essence an UML package diagram separating our technologies in 2 distinct groups, semantic web technologies and XML. The technologies are used to satisfy our requirements as they are shown in grey boxes in our previous requirements diagram

14
For our assessment we focus on rectify case as defined in GDPR

This crud matrix table tries to show that the rectify GDPR action is an umbrella case for other GDPR-defined actions.

Create, read, update and delete correspond to the listed GDPR terms, with modify also refers to modifying not just the flow or the data, but also access rights, metadata, etc. To be thorough, there is also another GDPR define action - 'Object to procedure', but it is excluded from this matrix because it is out of the system's boundaries.

15
This a base for our framework, BPMN model of the rectification process. It shows 3 swimlanes - each representing an actor - data subject, data controller and the third party. It shows actions, decisions and message exchange between the parties. Full resolution model available for discussion.

16
Moving on to our evaluation, as mentioned earlier we try to consolidate different technologies, that do have several things in common, but might use different approach, methodology or terms. It is a challenge to map those terms so they would all have some kind of common ground

So what we call a separation of concerns - a method that first uniquely identifies each technology and the terms and concepts it is using; and finally brings them together by finding the appropriate common denominator.

17
This could be probably easier to explain by looking at the table, where we define 4 main components or entities of our model: Message, Decision, Action and Flow.

In BPMN world messages are events, they can be defined either as triggers or reactions to another events. As far as ODRL is concerned, they translate into Actions, a native ODRL concept.

Decisions are gateways in BPMN, we utilise inclusive and parallel decisions, rules in ODRL

As far as schematron goes, as it is a schema language, the mapping is done through the predefined concepts that can be applied to the elements; elament body is applied to our message, rule to a decision and so on. Schematron is very simplistic in essence yet powerful

18
To validate our model, we have several options, but can divide the types of validations to 2 distinct groups: static and dynamic

In first case, we validate the syntax and structure through XML schema, as well as semantic web rules set by our DPRL profile together with SHACL constraints

In second case, and that is one of the main contributions of this research, following the idea of open, on-demand contract we try to dynamically validate the business logic - more specifically looking into the process flow and rules that are set. And for that we use schematron

19
To quickly introduce Schematron, it is one of the 5 XML schema languages; together with RELAX NG and NVDL it is a part of ISO DSDL (Document Schema Definition Languages) standard. There are 2 types of schema languages: grammar based and rule based.

A grammar-based schema language specifies the structure and contents of elements and attributes in an XML instance document. For example, a grammar-based schema language can specify the presence and order of elements in an XML instance document, the number of occurrences of each element, and the contents and datatype of each element and attribute.

A rule-based schema language specifies the relationships that must hold between the elements and attributes in an XML instance document. For example, a rule-based schema language can specify that the value of certain elements must conform to a rule or algorithm.

Schematron therefore, as a rule-based language, can be used to define co-constraints between elements fully and dynamically validating the business logic.

To constrain the structure, form, or syntax of XML instance documents, use a grammar-based schema language.

To specify data relationships, use a rule-based schema language

NVDL (Namespace-based Validation Dispatching Language) glues them together, creating one file validation strategy

20
A co-constraint is a dependency between data within an XML document or across XML documents.
Cardinality refers to the presence or absence of data.
An algorithmic check determines data validity by performing an algorithm on the data.