| | |
|---|---|
| Related to other papers in this special issue | 5 (p47); 20 (p199) |
| Addressing FAIR principles | A1.2, R1.1 (R1.2) |

# Ontology-based Access Control for FAIR Data

**Christopher Brewster[1,2†], Barry Nouwt[1], Stephan Raaijmakers[1] & Jack Verhoosel[1]**

[1]Data Science Department at TNO, Kampweg 55, Soesterberg 3769 DE, The Netherlands

[2]Institute of Data Science, Maastricht University, Maastricht 6229 ER, The Netherlands

**Keywords:** Access control; Semantic technology; Ontology; Security; FAIR

Citation: C. Brewster, B. Nouwt, S. Raaijmakers & J. Verhoosel. Ontology-based access control for FAIR data. Data Intelligence 2(2020), 66–77. doi: 10.1162/dint_a_00029

## ABSTRACT

This paper focuses on fine-grained, secure access to FAIR data, for which we propose ontology-based data access policies. These policies take into account both the FAIR aspects of the data relevant to access (such as provenance and licence), expressed as metadata, and additional metadata describing users. With this tripartite approach (data, associated metadata expressing FAIR information, and additional metadata about users), secure and controlled access to object data can be obtained. This yields a security dimension to the "A" (accessible) in FAIR, which is clearly needed in domains like security and intelligence. These domains need data to be shared under tight controls, with widely varying individual access rights. In this paper, we propose an approach called Ontology-Based Access Control (OBAC), which utilizes concepts and relations from a data set's domain ontology. We argue that ontology-based access policies contribute to data reusability and can be reconciled with privacy-aware data access policies. We illustrate our OBAC approach through a proof-of-concept and propose that OBAC to be adopted as a best practice for access management of FAIR data.

† Corresponding author: Christopher Brewster (E-mail: christopher.brewster@tno.nl, ORCID: 0000-0001-6594-9178).

## 1. INTRODUCTION

This paper focuses on the "accessible" aspect of the FAIR principles, arguing for the use of ontologies for fine-grained and secure access to FAIR data. In the twenty-five plus years since the creation of the World Wide Web, we have moved from a Web of documents to a vision for a Web of data [1, 2], especially open data and linked data [3]. The FAIR data principles represent the latest incarnation of a growing understanding of the implications of sharing data with other parties. While the "open data" movement represented a slightly utopian vision of the many possibilities of data sharing, the FAIR principles respect the reality of practical data sharing across different parties. For instance, the Accessible ("A") dimension of FAIR data can be seen as a parameter that varies between very open to very closed. Mons et al. [4] underline that "The "'A' in FAIR stands for 'Accessible under well-defined conditions'" and mention personal privacy, national security and competitiveness as reasons for data to be shielded.

The original home territory of the FAIR principles was life science research, where there is both a strong tradition of open databases and shared metadata principles [5], as well as a strong awareness of the sensitivity of certain types of data (like health or commercial data). Similarly, the law enforcement and security domains have long struggled with the need for "total information awareness" [6] (i.e., a philosophically similar concept as "open data") *versus* a "need to know" perspective on data access and use. For some actors such as police forces access to certain types of data is strictly controlled depending on authorization, warrants and court permissions and if rules are not strictly followed court cases can founder on the technicality of the evidence acquisition processes.

The FAIR principles have, to the best of our knowledge, not been applied routinely to data originating from sensitive sources, or data subject to confidentiality, privacy concerns, and security-related sharing restrictions. FAIR principles usually are applied to publicly available, open data, or carefully defined academic research data sets. And yet, law enforcement agencies (LEAs) typically gather data from a multitude of sources (including public domain data), and after further annotation, linking and enrichment, new confidentiality issues may arise. The resulting (meta)data would still demand data sharing and provenance tracking, but in a much more restricted fashion (see e.g. [7] for discussion of the data-centric world of law enforcement agencies and resulting issues with data sharing).

To this end, in this paper, we address FAIR principle A1.2: *The [accessibility] protocol allows for an authentication and authorization procedure, where necessary*. We propose to use ontology-based protocols for accessing object data through metadata. This allows for fine-tuned access, which additionally contributes to FAIR principle R1.1: *(Meta)data are released with a clear and accessible data usage license.*

Our primary goal is to utilize the concepts and relations from a domain ontology for data access control, in settings where data sharing is demanded but restricted. We propose Ontology-Based Access Control (OBAC) policies for accessing data on the basis of assigned metadata. OBAC is a new method that utilizes Semantic Web technologies and allows security policies to describe the *information* someone has access to irrespective of where this information resides. While not essential, an ontology based approach allows for easy integration with such FAIR principles as having unique identifiers (such as URIs) (F1), using open

and free protocols (A1) and guaranteeing a high level of conformity with the Interoperable principles (I1-3). Prerequisites for OBAC are:

1). Metadata needs to be assigned to raw object data prior to access.
2). The metadata scheme adheres to an ontology: it is hierarchically structured, with meaningful (interpretable, semantic) relations between nodes (concepts) and reflects domain knowledge.
3). Access to the object data occurs through the metadata, with the possibility of defining access for a given person or role to specific layers (strata) in the metadata.
4). Access to object data and re-usability of (meta)data is determined by referring to the structure of the metadata graph, the contents of the nodes, or both.

OBAC policies abstract from implementation details (e.g., resources like files or APIs) and thus are easier to reuse for new applications, or for integrating multiple information systems. The policies determine the information to which a user has (or has not) access to, based on the role of the user.

With OBAC, multiple resources can be captured in a single access policy. This makes the specification and maintenance of security policies more systematic and easier to manage. A single policy could, for example, give a user access to all pieces of information of a certain type, instead of writing out a policy for every piece of information separately. OBAC further allows for policies that enable access to data that comply with certain patterns in the implemented ontology. With the underlying data represented in an RDF triple store, these patterns express access to a single specific triple[①]. Thus, access can be defined at a very fine-grained level, which contrasts with traditional, high level access policies that use "all-or-nothing" approaches, like user name/password mechanisms. In the next section, we apply the OBAC approach to a use case from the security domain.

## 2. USE CASE

In many law enforcement scenarios, regulated data access is of prime importance. Similar to a house search warrant, police officers only can gather and inspect certain data on the basis of justified suspicion, and a mandate for pursuing their search. As an example, blindly going through the dark Web marketplace activity of a particular vendor of interest without an "information search warrant" will not yield evidence that would stand up in court. Here, OBAC is helpful. OBAC depends on a graph-based representation of metadata following Linked Data or knowledge graph principles. Usually, these metadata graphs are expressed as taxonomies (expressing strictly vertical, semantic relations between concepts) or ontologies that allow for modelling both vertical and horizontal semantic relations between nodes. Exploiting semantic relations between nodes for data access allows for "meaningful" access protocols.

① From a practical point of view, SPARQL-like graph patterns can easily be used to access the object data stored in the RDF triple stores.

An example of a metadata graph is depicted in Figure 1, which is a fragment of the categorization of illicit goods in the Agora dark Web market, generated automatically from a public Agora data dump [8][2].

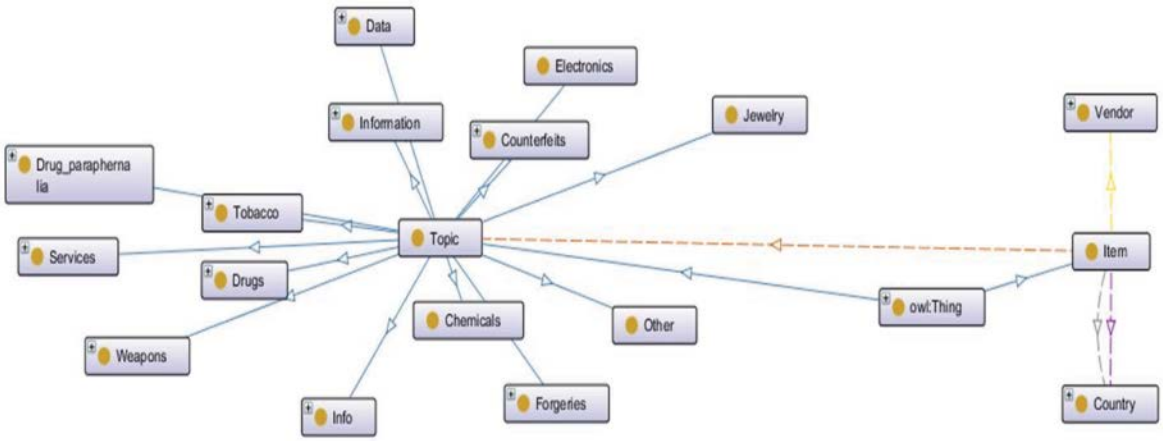


**Figure 1.** Categorization of illicit goods in the Agora marketplace.

Once raw data become linked to metadata graphs, structure-sensitive access protocols can effectively regulate views on the underlying raw data through traversal of the metadata graph. Possible restrictions involve semantic (ontology) labels on visiting nodes ("only party drugs are visible"), or topological aspects of the graph (such as the radius around nodes in the metadata graph that can be inspected by a given user). Even combinations of semantic and structure-based access restrictions are possible ("only party drugs statistics are visible, but not individual posts with vendor nicknames"). LEAs often start investigations with hunches. Information access protocols, backed by increasingly liberal investigation mandates, allow investigators to follow up on those hunches, by digging deeper in the data graph. We propose that OBAC is just the right approach to implementing such role- and context-dependent policies.

Figure 2 depicts access to a metadata graph, where every access level "sees" the information of the level above it, and at its own stratum. Starting from the top level ("Drugs", "Weapons"), with a metadata structure like "Drugs/ecstasy", "Drugs/cocaine", one can descend down the hierarchy, and finally (at the lowest level) obtain access to raw object data.

[2] Agora was a (now defunct) darknet marketplace for illicit goods *cf.* https://en.wikipedia.org/wiki/Agora_(online_marketplace).

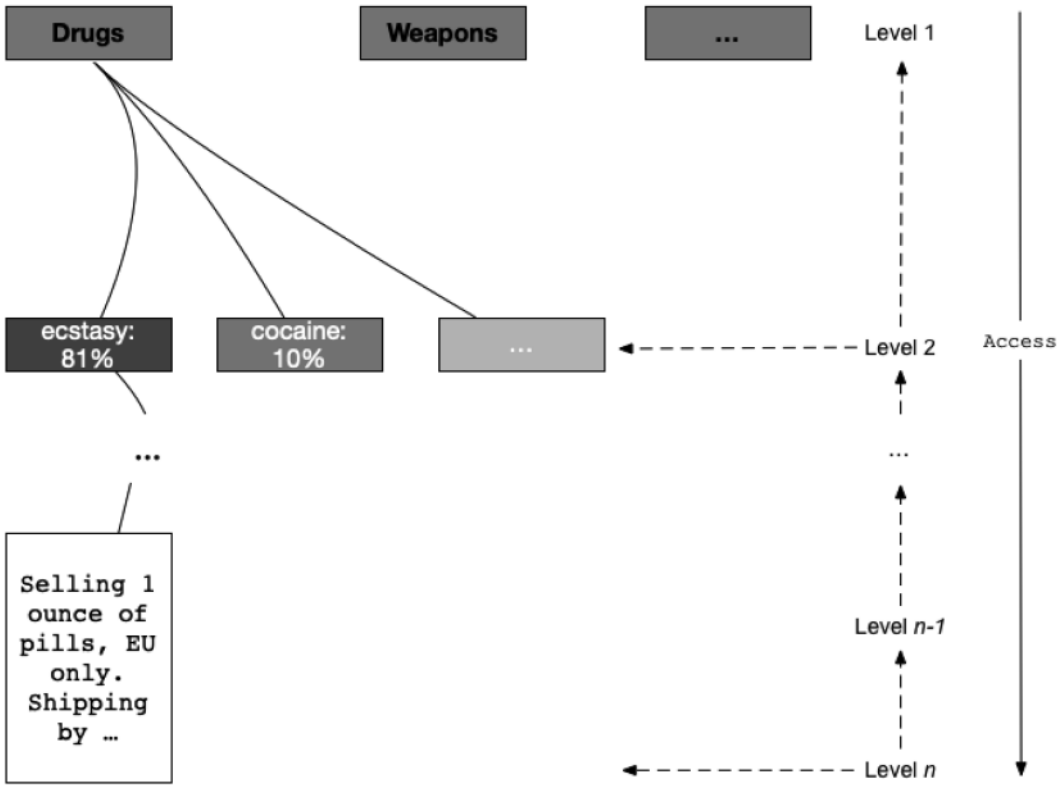


**Figure 2.** Access to object data through structured metadata.

## 3. PROOF OF CONCEPT

We implemented our OBAC approach by combining the open source SPARQL server Apache Jena Fuseki and the AuthzForce implementation of XACML. Apache Jena Fuseki (part of the Apache Jena framework) offers a SPARQL endpoint for accessing an RDF graph database [9]. XACML is a standard under development since 2001 by OASIS [10] for attribute-based access control usually expressed in XML (although now also in JSON) providing a means among other things for the description of security policies. AuthzForce implements the XACML standard [11] and consists of an authorization policy engine and a RESTful authorization server. Technical details can be found in the code on Github[3]. We utilize Apache Jena's extensible permissions layer[4] to intercept user SPARQL queries and send an appropriate decision request to AuthzForce's Policy Decision Point (PDP). The PDP evaluates the decision request against XACML authorization policies. Furthermore, we have extended AuthzForce with our custom *GraphPatternValue* and

[3] https://github.com/stephanraaijmakers/obac-fair-data.
[4] https://jena.apache.org/documentation/permissions/index.html.

*GraphPatternValueMatchFunction* to allow XACML security policies to include SPARQL-like graph patterns and let AuthzForce match those to parts of the user SPARQL query and decide whether access is approved or rejected.

Our proof of concept is based on the following, typical LEA scenario:

> *Two narcotics police officers, one from the United States and one from Australia, receive a limited mandate to investigate a dark Web marketplace, but only for narcotics that both originate from and are shipped to their own country. The limited mandate only allows them to look at the title, origin and destination of items. The narcotics officers use their mandate to browse the data and collect facts that might be worth further investigation. If approved, they will be allowed to work with a broader mandate. The broader mandate allows them to hone in on a more specific class of drugs (Steroids) in the data set.*

We applied our approach to the aforementioned Agora dark Web marketplace data set, a CSV file with over 100,000 unique listings (rows) and 12 columns describing listings (Vendor, Category, Item, Item Description, Price, Origin, Destination, Rating and Remarks). We took the first 5,000 records from this data set and linked them to the Agora ontology shown above in Figure 1. Columns of the data set are mapped to the ontology in the following way; every row (i.e., marketplace item) becomes an individual of the Item class with the following properties.

Object property assertions:

- hasDestination: the destination country of the item.
- hasOrigin: the origin country of the item.
- hasVendor: the vendor of the item.
- hasTopic: the topic (category) of the item.

Data property assertions:

- hasRating: the rating of the item.
- hasPrice: the price of the item.
- hasDescription: the description of the item.
- hasRemark: the remark of the item.
- hasTitle: the title of the item.

This mapping allows us to convert the CSV Agora data set into RDF data using the OpenRefine[5] tool. Although OBAC does not restrict the location and format of the data under its control (i.e., its security policies only describe *what* information someone has access to in terms of the domain ontology), we transformed the Agora data set into a single RDF graph to simplify the experiment. We load this RDF data into an Apache Jena triple store and used the following security policies.

[5] http://openrefine.org/.

The security policies for this scenario involve the following four roles:

- USA_drugs_officer_limited: the role for a drugs officer from the United States with a limited mandate.
- AUS_drugs_officer_limited: the role for a drugs officer from Australia with a limited mandate.
- USA_drugs_officer_broad: the role for a drugs officer from the United States with a broad mandate.
- AUS_drugs_officer_broad: the role for a drugs officer from Australia with a broad mandate.

One of the graph patterns for the role *USA_drugs_officer_limited* looks like this[®]:

?item **dwo:hasTitle** ?object .
*IF*
?item rdf:type dwo:Item .
?item dwo:hasDestination **dwd:USA** .
?item dwo:hasOrigin **dwd:USA** .
?item dwo:hasTopic ?topic .
?topic rdf:type ?topicType .
?topicType rdfs:subClassOf+ **dwo:Drugs** .

This graph pattern should be read as:

This role gives access to the title of an item, if the item has USA as its destination and origin country, and is drugs related.

The above policy will be repeated for the other properties apart from "title" to which this role should have access. Once these security policies are in place, the USA drugs officer can inspect the data set safely without risking trespassing the mandate. The officer might be interested in all items he has access to and composes the following SPARQL query:

```
  SELECT ?item ?title ?topic
WHERE {
  ?item a dwo:Item .
  ?item dwo:hasTitle ?title .
  ?item dwo:hasTopic ?topic .
}
```

® For readability we included prefixes (where the prefix *dwo* stands for **d**ark **w**eb **o**ntology and *dwd* for **d**ark **w**eb **d**ata) and the IF keyword; these are not part of the actual implementation.

The above query gives the following example results (Table 1).

**Table 1.** Example results of SPARQL queries.

| Item | Title | Topic |
|---|---|---|
| dwd:item/124 | 4-FMA...!! 1 KG | dwo:Drugs/RCs |
| dwd:item/2423 | Ephedrine HCL 8mg (50x Pills 400mg Total) | dwo:Drugs/Weight_loss |
| dwd:item/2561 | ORBIS PHARMA LABS - TEST C 10ML VIAL X 1 - Testosterone Cypionate - $67 | dwo:Drugs/Steroids |

Note: These results only include items that have USA as their origin and destination, even though this is not explicitly mentioned in the SPARQL query.

From this data browsing session, the officer finds hints that especially the Steroids related drugs are increasing in the USA region. To further investigate this, the officer gets another role with a broader mandate that is needed to access vendor information. The graph pattern for the role *USA_drugs_officer_broad* that reflects this broader mandate is as follows:

?item **?predicate** ?object .
*IF*
?item rdf:type dwo:Item .
?item dwo:hasDestination **dwd:USA** .
?item dwo:hasOrigin **dwd:USA** .
?item dwo:hasTopic ?topic .
?topic rdf:type ?topicType .
?topicType rdfs:subClassOf+ **dwo:Steroids** .

This graph pattern can be read as:

This role gives access to all properties of an item, if the item has USA as its destination and origin country, and its topic is steroids related.

With the broader mandate the officer can follow up, and submit the following query:

```
  SELECT ?item ?title ?vendor
WHERE {
  ?item a dwo:Item .
  ?item dwo:hasTitle ?title .
  ?item dwo:hasVendor ?vendor .
}
```

This query produces the following result (Table 2).

**Table 2.** Example results of SPARQL queries.

| Item | Title | Vendor |
|---|---|---|
| dwd:item/2561 | ORBIS PHARMA LABS - TEST C 10ML VIAL X 1 - Testosterone Cypionate - $67 | dwd:InsideTheWhale |
| dwd:item/2563 | Nordicor Trenbolone Enanthate 200mg/ml 10 vial | dwd:pharmacy_land |
| dwd:item/2578 | Rip 300 10ml multi injection vial | dwd:cerberus |

Table 2 lists user specific (vendor) names. This demonstrates that the use of OBAC security policies as metadata to the Agora data set, makes the Agora data set more FAIR by providing an explicit and machine readable description of the allowed access and show that our proof-of-concept implementation can enforce them.

## 4. RELATED WORK

Attribute based access control (ABAC) has become the standard approach to providing differential access to data in databases [12, 13]. There exists a body of work using ontologies and semantic principles to simplify and systematize the ABAC approach. Reference [14] focuses on using ontologies to simplify XACML, while [15] and [16] both use OWL to provide access control mechanisms. OBAC differs from these approaches in that it does not use an ontology to describe the users, roles and permissions, but it uses the domain ontology to specify the exact resources (or information) a particular user has access to or not. Reference [17] is an example where the ontology is used for both these purposes and this is also part of the future work for OBAC. Furthermore, the other approaches focus on access control to Web services, while our approach is independent of the location of the data. Reference [18] proposes OBDA, Ontology-Based Data Access, which emphasizes data consistency through metadata schemes and data content. Our OBAC approach is complementary to OBDA.

Previous work on the application of the FAIR principles in the context of restricted access to data has been limited. Seemingly operating under a misunderstanding of the FAIR principles, the research in [19] proposes an "extension" to the existing FAIR principles to specifically accommodate the needs of sensitive clinical data where privacy needs to be maximized. The authors' approach is particularly influenced by the requirements of the GDPR. In a similar spirit, the authors in [20] propose defining specific FAIR data types to capture privacy aspects in the metadata together with a coherent data access governance framework. More generally, the OMETA system of [21] exploits metadata for the configuration, capture, inspection and distribution of biomedical data.

## 5. CONCLUSION AND FUTURE WORK

We have outlined a technical approach to applying ontology-based authentication protocols to structured data. Exploiting the structure of the applied ontology, our approach allows for data access control, by assigning semantic or topological restrictions to users, based on individual or group-based policies. Our work targets the FAIR principles A1.2 and R1.1, and thus affects both data access authentication and data usage. In interactive scenarios, where data users can modify or assign metadata themselves, our methods also can be extended to fine-grained metadata provenance control (FAIR principle R1.2). The approach we propose is entirely complementary with Landi et al. [22] and provides a way to operationalize the requisite access controls. The policy-based approach we espouse can be adapted to fit with legal requirements such as GDPR or even the protection of copyright in some circumstances [23].

The OBAC approach is currently being evaluated in on-going research to enable European LEA's to inspect dark Web data. We intend to apply OBAC to other domains as well, including health. Future work will address systematic descriptions and implementations of different ontology-based access policies, including ones that restrict information access based on navigational limits, or more semantics-oriented policies described by an ontology. Furthermore, we need to model users using ontologies thereby combining both role-based and attribute-based access control methods. Adding methods for user interaction with implemented policies, like negotiation-based policy bypassing (e.g., trading one restriction in for another) or quota-based data access will make OBAC more user-oriented.

## AUTHOR CONTRIBUTIONS

The authors worked collaboratively on the structure and outline of the paper. S. Rraaijmakers (stephan.raaijmakers@tno.nl) initiated the paper and made the initial connection of OBAC with the security domain. B. Nouwt (barry.nouwt@tno.nl) and J. Verhoosel (jack.verhoosel@tno.nl) developed the OBAC core ideas, while B. Nouwt implemented the approach and provided the PoC description. S. Rraaijmakers and C. Brewster (Christopher.Brewster@tno.nl) worked on the introduction, conclusions and related work and were responsible for overall editing.

## ACKNOWLEDGEMENTS

Part of this work was supported by the Titanium Project (funded by the European Comission under grant agreement 740558)[7]. The work was also supported by TNO's internal research project "ERP AI".

[7] https://www.titanium-project.eu/.

## REFERENCES


[1] T. Berners-Lee. Information management: A proposal. (1990). Available at: https://www.w3.org/History/1989/proposal.html.

[2] T. Berners-Lee, J. Hendler & O. Lassila. The semantic web. Scientific American 284(5) (2001), 34–43. Available at: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.

[3] C. Bizer, T. Heath & T. Berners-Lee. Linked data – The story so far. International Journal on Semantic Web and Information Systems 5(3)(2009), 1–22. doi:10.4018/jswis.2009081901.

[4] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L.O. Bonino da Silva Santos & M.D. Wilkinson. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use 37 (2017), 49–56. doi:10.3233/ISU-170824.

[5] M.D. Wilkinson, M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, … & B. Mons. The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3(2016), 160018. doi: 10.1038/sdata.2016.18.

[6] W. Safire. You are a suspect, The New York Times. (2002). Available at: https://www.nytimes.com/2002/11/14/opinion/you-are-a-suspect.html.

[7] S. Wood. The paradox of police data. KULA: Knowledge Creation, Dissemination, and Preservation Studies 2(2018), 9. doi:10.5334/kula.34.

[8] P. James. Dark net marketplace data (Agora 2014-2015). (2017). Available at: https://kaggle.com/philipjames11/dark-net-marketplace-drug-data-agora-20142015.

[9] Apache Jena – Apache Jena Fuseki, Apache, 2018. Available at: https://jena.apache.org/documentation/fuseki2/.

[10] B. Parducci, H. Lochhart & R. Levinson (eds.) OASIS eXtensible access control Markup Language (XACML) TC. (2017). Available at: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml.

[11] AuthZForce (Community Edition) – Authzforce, 2019. Available at: https://authzforce.ow2.org/.

[12] E. Yuan & J. Tong. Attributed based access control (ABAC) for Web services. In: IEEE International Conference on Web Services (ICWS'05), 2005, pp. 561–569. doi:10.1109/ICWS.2005.25.

[13] V.C. Hu, D. Ferraiolo, R. Kuhn, A. Schnitzer, K. Sandlin, R. Miller & K. Scarfone. Guide to attribute based access control (ABAC) definition and considerations, National Institute of Standards and Technology, 2014. doi:10.6028/NIST.SP.800-162.

[14] T. Priebe, W. Dobmeier & N. Kamprath. Supporting attribute-based access control with ontologies. In: The First International Conference on Availability, Reliability and Security (ARES'06), 2006, pp. 465–472. doi:10.1109/ARES.2006.127.

[15] H. Shen. A semantic-aware attribute-based access control model for Web services. In: A. Hua & S.-L. Chang (eds.) Algorithms and Architectures for Parallel Processing. Berlin: Springer, 2009, pp. 693–703.

[16] N.K. Sharma & A. Joshi. Representing attribute based access control policies in OWL. In: 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), 2016, pp. 333–336. doi:10.1109/ICSC.2016.16.

[17] A. Padia, T. Finin & A. Joshi. Attribute-based fine grained access control for triple stores. In: The 3rd Society, Privacy and the Semantic Web – Policy and Technology Workshop, the 14th International Semantic Web Conference, 2015. Available at: https://ebiquity.umbc.edu/paper/abstract/id/706/Attribute-based-Fine-Grained-Access-Control-for-Triple-Stores.

[18] M. Console & M. Lenzerini. Data quality in ontology-based data access: The case of consistency. In: The Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014, pp. 1020–1026. Available at: https://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8552.

[19] P. Holub, F. Kohlmayer, F. Prasser, M.Th. Mayrhofer, I. Schlünder, G.M. Martin,… & J.-E. Litton. Enhancing reuse of data and biological material in medical research: From FAIR to FAIR-Health. Biopreservation and Biobanking 16(2)(2018), 97–105. doi:10.1089/bio.2017.0110.

[20] M. Corpas, N.V. Kovalevskaya, A. McMurray & F.G.G. Nielsen. A FAIR guide for data providers to maximize sharing of human genomic data. PLOS Computational Biology 14 (2018), e1005873. doi:10.1371/journal.pcbi.1005873.

[21] I. Singh, M. Kuscuoglu, D.M. Harkins, G. Sutton, D.E. Fouts & K.E. Nelson. OMeta: An ontology-based, data-driven metadata tracking system. BMC Bioinformatics 20(2019), 8. doi:10.1186/s12859-018-2580-9.

[22] A. Landi, M. Thompson, V. Giannuzzi, F. Bonifazi, I. Labastida, L.O. Bonino da Silva Santos & M. Roos. The "A" of FAIR – as open as possible, as closed as necessary. Data Intelligence 2(2020), 47–55. doi: 10.1162/dint_a_00027.

[23] I. Labastida & T. Margoni. Licensing FAIR data for reuse. Data Intelligence 2(2020), 199–207. doi: 10.1162/dint_a_00042.