On the way to measure KG transparency: Formalizing transparency - Requirements and first models

Jennie Andersen, Sylvie Cazalens, Philippe Lamarre

On the way to measure KG transparency: formalizing transparency - requirements and first models.

ANR Project DeKaloG*
Research report D3.1

Jennie Andersen, Sylvie Cazalens, and Philippe Lamarre
INSA Lyon, CNRS, LIRIS UMR5205, F-69621 France

December 15, 2021

1 Introduction

The demand for transparency is high and increases in many domains including everyday life (politics, government, administration, business, environment, health, personal data management...) and science. Although very often cited and desired, the contours of this notion often remain rather vague and can even differ significantly from one context to another. Such lack of precision facilitates the possibility for an element to be qualified as transparent by its promoters without this element actually meeting the expectations of outside observers. For example, many governments have made significant efforts to make open data available. However, from the public's point of view, there are limitations that can hinder transparency: file formats that limit queries (simple scans), data formats that are difficult to use, little information on how certain data are produced, difficulty integrating with other sources, etc. The use of norms and standards for data publication (domain modeling, data format, access and query protocols...) brings an obvious improvement. In particular, Knowledge Graphs make it possible to represent, publish, make accessible and searchable information in any field. Whatever the considered interpretation of transparency, the use of a public KG constitutes a major element that can only improve transparency. However, using a KG is not in itself sufficient to fully guarantee transparency.

In this report, our first objective is therefore to better understand and formally define the concept of transparency. First, in section 2, we present a review of the different definitions of transparency that can be found in the literature of many domains, as well as many of its companion concepts that are very often used to complement and to clarify it: accessibility, accountability, findability, interoperability, provenance, privacy, quality, verifiability, replicability, repeatability, reproducibility, understandability, trustworthiness, openness, completeness, etc. The synthesis makes us distinguish between notions that characterize data sources and those that characterize observation needs and leads us to pose our own informal definition of transparency. Then, in section 3, we propose an formal definition of the concept of transparency in a general context. In section 4, we adapt it to knowledge graphs. Finally, section 5 presents first elements about measures.

*AAPG2019 - ANR Project - Funding Instrument PRC, CE3 - Intelligence Artificielle.

2 Different transparency interpretations and companion concepts

2.1 Transparency : a contextual meaning

Although transparency is cited in many papers, it remains a rather ill-defined concept when going beyond the general definition of a dictionary, with no uniform view of what constitutes it [MJ15]. This section does not aim at an exhaustive listing of the different meanings of transparency in the literature. Rather, through the choice of some representative papers, we analyze several definitions and uses.

Starting from a general context, we first pick up the definition of the word *transparent* by the Merriam-Webster: "characterized by visibility or accessibility of information especially concerning business practices" [Mer16].

Visibility and accessibility are also mentioned in the context of governance, politics and laws in [Wya18] where several definitions of transparency are provided. They range from "a metaphorical window [...] allowing outsiders of [...] a decision-making body to gain [...] visual access to that body's insides." to others "based on the availability of information". Some others describe transparency slightly differently: "in a relational manner", between observer and observed or ruler and ruled.

The fact that several parties are involved also appears for transparency of governments and the provisioning of information. In [MJ15], which context is big and open linked data, the authors underline that "Transparency is aimed at overcoming the information asymmetry between the government and the public." In the same spirit, [SLC20] propose the following definition of transparency: "The availability of information about an actor that allows other actors to monitor the workings or performance of the first actor." that explicitly introduces monitoring.

This possibility of monitoring, checking, verifying appears in many domains requiring transparency, such as journalism, with the development of fact checking. This latter domain especially encourages journalists to be more transparent about their methodology and sources to increase their trustworthiness: "transparency refers to a verification process presented in a way that allows the audience or readers to decide for themselves why they should trust or distrust it" [BFC18]. As another example, biology scholars developed their own platform Galaxy[1] in order to promote transparency, as well as accessibility and reproducibility [GNT10], offering not only the possibility to check results but also to reuse them.

Considering Computer Science, the common usage of transparency is rather recent and is growing with the use of big data along with personal data and machine-learning [Ber+19]. Transparency of digital objects, from data to softwares, becomes a key point. For instance, [Ber20] introduces several definitions: "the ability to access and work with data no matter where it is located" and "the guarantee that the data being provided is accurate and from some official source." She considers that these definitions are no longer adequate so she proposes a new one, privacy-oriented: "Data transparency is the ability of subjects to effectively gain access to all information related to data used in processes and decisions that affect the subjects." Transparency as a dimension of the quality of the data is defined as follows: "Transparency is the ability to interpret the information extraction process in order to verify which aspects of the data determine

[1] `www.galaxyproject.org`

its results." [FTT19]. And it may also be required for algorithms: "An algorithm is transparent (for a specific purpose) if it discloses its motivation and actions." [ABV18].

These selected examples already show that the definitions of transparency are very contextual. Obviously, they depend on the subject to be transparent: transparency of a government is not the same as transparency of data. But even for the same object, there can be significant variations, as shown by [Wya18] for government and by [Ber20] for data. But all definitions aim at provisioning more information. They require the access to information, utilizations and motivations related to the data and process.

Moreover, transparency can only be truly attested by an external observer (subjects, the public...). Hence, transparency requires an observed element, an observer [SLC20] and a mean of observation [Wya18]. Finally, as mentioned in the last definition, transparency is needed "for a specific purpose" which is user defined.

2.2 Related concepts

Many concepts are associated with transparency. They can be used to describe situations where the need for transparency is important or to specify how to obtain it. In any case, studying them can only enhance our understanding. Without claiming to be exhaustive, we present here those that seem to us to be the most frequently associated with the concept of transparency in different fields. For each of them, we propose to retain an intuitive definition in order to obtain a set of coherently defined concepts.

2.2.1 Openness

It corresponds to the idea of making available, open, all data or information one has. This idea is especially present in the field of public administration. Indeed, asking that all the information used by an administration be made public is quite an intuitive solution to bridge the level of information gap between on one side the administration and on the other side the public. Such consideration has paved the way for the emergence of open data [HV11] which also offers the opportunity to reuse data in order to extend knowledge and create new quality products and services. In some cases, transparency and openness are so close that they are considered as synonymous [Wya18; MJ15]. Otherwise, the limit between these two is unclear when they are defined together by "There should be no secret record keeping. This includes both the publication of existence of such collections, as well as their contents." in [SLC20]. Indeed, if the actors who require transparency share the same goals as those who open the data they have collected, hiding nothing should allow both to work properly. A difference between the objectives of one and the other induces a divergence between these concepts quite easy to observe. For instance, [MJ15] shows that while openness is a necessary condition for transparency, only accountability (see after) can testify if the data are sufficient to be transparent, if enough information has been recorded and disclosed.

In order to take into account this possible need for additional information, the intuitive definition that we propose to retain voluntarily and explicitly limits the scope of openness to available information.

Intuitive definition 1.
Openness is the full disclosure of all available data and information.

2.2.2 Accessibility

While openness is necessary for transparency, it is not enough. Indeed, many of the previous definitions of transparency also explicitly require that the information be accessible, in the sense of being reached, used or seen. For instance, accessibility directly appears in Bertino's definitions such as "effectively gain access to all information" and "the ability to access data" [Ber20].

In the domain of data management, accessibility is one of the FAIR principles [Wil+16], which have been proposed in order to improve machine actionability. In particular, they aim at acting "as a guideline for those wishing to enhance the reusability of their data holdings", from data to algorithms and workflows. The FAIR principles are significant but partly out of our scope, mainly because transparency does not rely on reusability. This is why we only keep the part on accessibility which requires that: "A1. (meta)data are retrievable by their identifier using a standardized communications protocol ; A1.1 the protocol is open, free, and universally implementable ; A1.2 the protocol allows for an authentication and authorization procedure, where necessary ; A2. metadata are accessible, even when the data are no longer available".

Finally, accessibility also appears as a dimension of quality in knowledge graphs (KG). In [Fär+18], accessibility is "the extent to which data are available or easily and quickly retrievable", and it can be measured according to seven criteria: "dereferencing possibility of resources, availability of the KG, provisioning of public SPARQL endpoint, provisioning of an RDF export, support of content negotiation, linking HTML sites to RDF serializations, provisioning of KG metadata". It is also usually linked with a need of licensing [Fär+18; Zav+16; HP20] but also interlinking [Fär+18; Zav+16] and sometimes of security and performance [Zav+16]. Whereas in [HP20], "Accessibility implies that data or part of it must be available, retrievable, and contain a license.", we adopt a more general definition, which does not consider the question of license, points A1 and A2 of the FAIR principles being left implicit in the case of digital assets [Fär+18].

Intuitive definition 2.
Accessibility is "the extent to which data are available or easily and quickly retrievable".

2.2.3 Accountability

Accountability is a major topic in political and governmental research about transparency. For instance, transparency of governments in the Big and Open Linked Data in [MJ15] is described as being synonymous of both openness and accountability. In this paper, "accountability implies answerability for one's actions or inactions and the responsibility for their consequences ; accountability means also taken responsibility for decisions."

In Computer Science, accountability also appears closely related with transparency such as in: "Information accountability means the use of information should be transparent so it is possible to determine whether a particular use is appropriate under a given set of rules and that the system enables individuals and institutions to be held accountable for misuse." [Wei+08]. In [OH20], a new metadata model, LiQuID, is defined to make datasets accountable throughout their lifecycle: "Accountable datasets are datasets about which there is sufficient information to justify and explain the actions on these datasets to a forum of persons, in addition to descriptive information and information on the people responsible for it.". In other words, transparency of relevant data is required at all stages of the lifecycle.

In all these works, accountability requires transparency and goes beyond the general need of more information by specifying it: involving, at varying degrees, people's responsibility and relevance and completeness of information in order to justify or verify data use/misuse. Hence, we adopt the following definition[OH20; Wei+08].

Intuitive definition 3.
Accountability means that "there is sufficient information to justify and explain the actions on these [data] to a forum of persons, in addition to descriptive information and information on the people responsible for it" ; "that the system enables individuals and institutions to be held accountable for misuse" of the data.

2.2.4 Provenance

According to [FTT19], in order to measure data quality, transparency can be divided into data provenance and explanation in order to "interpret the information extraction process" and "verify which aspects of the data determine its results". In their work, data provenance is based on "meta-data describing where the original data come from".

In Computer Science, this concept is rather well known. In the context of relational databases, provenance can be defined as extra information about how a query result comes from, or how it has been computed [Sen19]. Over the past twenty years, computing data provenance in database systems has been well studied, with some results that can be adapted to the context of knowledge graphs. Its computation however is not always easy. Research about provenance in scientific workflows systems [Bou+17; GSB16], and more generally scientific databases[GNT10], enables reproducibility, concept that is discussed in the next subsection.

In [SLC20], provenance is depicted as a central element of a knowledge graph. A statement is associated with provenance information saying who, where, when and how was it made. Among the example rules applying to a knowledge graph, one of them, a transparency rule explicitly requires provenance of the document at the origin of the focused statement.

In [OH20], provenance is a dimension of the LiQuID model. In fact, for each step of the data lifecycle, provenance information partly answers the questions who, when, where, how and what, with at least a description and possibly an explanation. This link is shown by comparing the LiQuID metadata model and the PROV metadata model [GM13]. The PROV data model is a recommendation from the W3C which aims at representing the provenance of digital objects, i.e. their origins, their modifications over time, the people who act on them and so on. They end with the following definition that we adopt [Mor+15]:

Intuitive definition 4.
"Provenance is a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing."

2.2.5 Reproducibility & co.

When speaking about transparency in sciences and especially in experimental sciences, one thinks of the way the experiments have been conducted, in which environment, etc. Indeed, "Science thrives on reproducibility" [Geo14]. And to be reproducible, a high level of transparency is required.

In fact, reproducibility has different definitions, and is usually understood with (at least) two other 'R' notions : *replicability* and *repeatability*. According to the ACM [ACM20], all three notions require to obtain similar results on multiple trials of an experiment. They differ in who and in what conditions the experiment is realized. Repeatability concerns experiments made by the same team in the same experimental setup ; reproducibility requires a different team in the same experimental setup ; and replicability involves a different team in different experimental setup. For the following, the term reproducibility implicitly refers to these three notions.

Hence, transparency is essential for reproducibility in many different fields, from psychology to computer sciences, including economics and many others. The need for reproducibility has been

increasing along with the studies about the reproducibility crisis. Thus, there is also a growing need of transparency on data, methods and research materials to enable this reproducibility [SR18]. In artificial intelligence [Hai+20] it takes the form of disclosing the computer code, the parameters and the dataset used, but also the whole data processing and training pipeline. In this case, reproducibility partly relies on provenance information. However, as discussed before, privacy concerns may reduce transparency, especially if the dataset holds personal data. Hence, reproducibility is also affected. In this work, we will consider the ACM definitions [ACM20].

Intuitive definition 5.
Reproducibility and replicability means that "the measurement can be obtained with stated precision by [the same or a different] team using [the same or a different] measuring system in [the same or a different] location on multiple trials."

2.2.6 Privacy

Bertino [Ber20] closely relates transparency with the notion of privacy. According to her definition, transparency is giving access to information, to data, but not just any: the data disclosed should expose, not the personal data itself, but the use of these data. Data transparency is then to give access to more information in order to support data privacy. This is captured in this definition: "Data transparency is the ability of subjects to effectively gain access to all information related to data used in processes and decisions that affect the subjects."

This link with transparency which supports data privacy can be surprising because more often, privacy is presented another way round where it stands against transparency. Indeed, privacy is also the "controlled access to information related to an agent" [SLC20]. In this case, privacy aims at restricting the access to data, and so being less transparent. The opposition between these two notions can be expressed as follows: "Privacy tends to limit or restrict actions over information items, whereas Transparency tends to allow (in some cases, mandate) actions over them, which explains the natural tension that exists between the two." [SLC20]. [Wei+08] makes the following distinctions between this two sides of privacy: Privacy as understood by Bertino is "privacy rights as they relate to the collection and use of personal information" is opposed to "other privacy protections that seek to preserve control over, say, one's physical integrity."

Still, these two visions may not be in contradiction, as Bertino does not require that all personal data are being divulged to everyone, and as the definition of privacy only imposes a controlled access, not a closed access.

Secrecy is a concept close to privacy, as it usually tends to limit the disclosure of information. It is also more general as it is not limited to personal data but can apply with companies and governments "government secrecy [is] the logical antonym of transparency" [Wya18].

For privacy rights, we adopt the following definition, extracted from the definition of data transparency from [Ber20].

Intuitive definition 6.
Privacy rights means that subjects are able to acquire "all information related to data used in processes and decisions that affect the subjects."

2.2.7 Explanation & Understandability

According to [FTT19], transparency can be divided into *data provenance* and *explanation* where explanation "describes how a result has been obtained". It seems evident that giving a good explanation increases transparency. But it is also very dependent on the one receiving it. Indeed, a useful explanation has to be adapted to the audience.

This lead us to consider the concept of understandability. It "emphasizes on explaining results or processes to make these transparent to some audience" [HDL17]. Understandability also increases transparency, but it is still very dependent on the audience.

In our research, we do not suppose that we know the audience, hence we consider that explanation and understandability are not directely in our scope. Moreover, this definition of explaining is partly included in our definition of provenance, the difference lies in how the description is made, if it explicitly requires that it is human-readable and human-understandable.

2.2.8 Verifiability and verification

The motivations leading a source to make information public are diverse. Whatever they are, in one way or another, the publication is expected to affect the readers' beliefs. The latter may, for example, accept the information as true or, on the contrary, consider it to be false (and, in the process, perhaps also lower their confidence in the source). Here, we only consider information that a source hopes the reader will accept. The verifiability of such information is an important element in this process because it is supposed to facilitate verification. Indeed, by propping up the information, whether it is carried out directly or by a trusted unbiased third party, verification paves the way for its acceptance: verifiability also enhances trustworthiness [Fär+18; Zav+16]. Here, different notions have to be distinguished: verification, verified and verifiability. Verification is carried out by an actor who uses all the means at his disposal. When the conclusion is positive, the information is deemed "verified by the agent". The verifiability of information is related to the elements available to an actor to help him conducting a verification. These notions are therefore very close but different and complementary. In the following, we adopt the definition below [Fär+18].

Intuitive definition 7.
"Verifiability is the degree and ease with which the data can be checked for correctness".

Accordingly, the transparency of a source consists in providing enough elements to enable a person to carry out an audit. Since we all have different abilities, skills and resources, the needs may differ from person to person. Note it is often expected that these correspond to the way in which the source has itself obtained the information, which brings provenance into the picture [Zav+16]. In the more specific context of scientific information, verifiability can be achieved through the reproducibility of experiments [GNT10].

For a person who is trying to make up his or her own opinion about a piece of information, but does not have the necessary resources to conduct an audit (even if only the time), the above elements may still be of some interest. However, as explained above, knowing that third parties have conducted audits, and having access to their comments and conclusions, is may be more important [Zav+16]. Journalists, fact-checkers and many others are aware of that. Hence, we retain the following definition [MB13].

Intuitive definition 8.
"A verified data is information that has been vetted by a third party."

A link with transparency is also present here. "A second attribute that increases the inferability [transparency necessary condition] of information is verification."[MB13]. In other terms, with complete transparency, a source should quote and reference the verifications carried out by third parties.

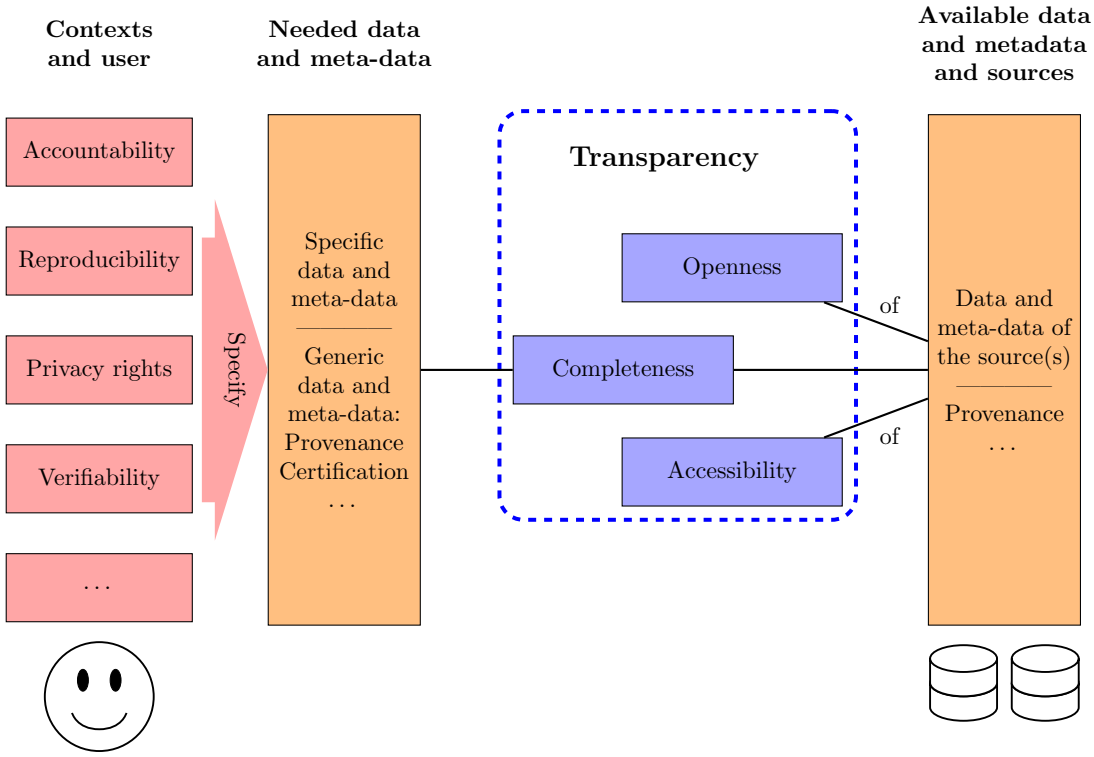
To conclude, measuring transparency for an audit purpose can be difficult. Indeed, it is relative both to the degree of the audit and to person conducting it. At the very least, clear

assumptions about the resources and skills assumed would be necessary to implement such measures. On the contrary, measuring the presence of references to third-party audits is easy, as long as one does not try to judge either their potential bias or the trust placed in them.

2.3 Synthesis

Our previous study shoes that various definitions of transparency exist, presenting some similarities: they require the access to more information (openness and accessibility) and underline that different types of actors are concerned, such as information providers and consumers. But, these definitions are also very contextual, showing in addition a diversity of the concepts they relate to transparency. Hence, in order to apply to any context, we propose the following definition of transparency.

Intuitive definition 9.
Transparency aims at giving access to the suitable information regarding the needs of observation to any observer.



Reproducibility is understood in the broad sens, with *repeatability* and *replicability*.
Privacy rights focuses on the part of *privacy* which require more information (not the one that aim at restricting access to information).

Figure 1: Transparency and companion concepts

This formulation combines the elements introduced by the different definitions of section 2.1. Observation is the first one. Indeed, whatever the definition considered, transparency tends to

satisfy a need for observation. If the shared requirement is the possibility of observation, the elements for which observation is required may vary greatly from one observer to another or from one context to another. In other terms, to understand a transparency request, the observer's need must therefore be specified. Finally, to obtain transparency, it is mandatory to provide, to give access to information suitable regarding the need: the information must meet the need in the most appropriate and complete way. For example, if the need expressed by the observer is a need of reproducibility of a numeric process, then transparency consists in giving, among other things, full access to the source code, the data used and the environment of execution. But, there is no interest in observing information with no link with the need, as the color of the computer or the weather broadcast: it is not suitable information.

To go further, figure 1 presents a whole picture, taking into account the related notions presented before. *Accountability*, *reproducibility*, *privacy rights*, *verifiability*... are different contexts. Each of these contexts specifies (observation) needs of data and meta-data, some of which are specific, others being more generic as for example provenance or certification information. On the other side, data sources record data and meta. Central to the picture is transparency, which relies not only on openness (increasing the data exposed) and accessibility (which garanties that one can read or work these data) but also on the completeness of the data provided by the source with respect to the data and meta-data forming the observation need. In the end, the question of transparency arises as a question of completeness between the information needed to meet the need and the information actually available in the sources: "Is the set of accessible information complete enough to satisfy the observation need ?"

In the following, we seek to define transparency more formally. We first do so independently of the type of sources. Then we focus on knowledge graphs, without any assumption on the application domain.

3 General formalization of transparency

The purpose of this section is to define transparency formally. Beyond openness and accessibility, transparency depends on the specification of the observation need and on the completeness of the available data. The way the former is defined impacts the latter.

3.1 Preliminary notions.

Several notions are involved when considering transparency, such as information or source of information. We use the following definitions, trying to be as precise as possible while remaining general.

3.1.1 Information item

We introduce information items as pieces of information. Nothing is assumed about the way they are stored and consulted, nor on their format. They can represent data as well as meta-data or knowledge.

Definition 3.1 (Information item)**.**
An ***information item****, denoted* i*, is a piece of information which can be stored, processed or communicated.*
The theoretical ***set containing all information*** *is noted* $\mathcal{I}$*.*

3.1.2 Source

We idealize a source as a set of information plus a support enabling it to be stored and interrogated. This is appropriate for a newspaper (the support can be paper or electronic), for a person (who is her own support) or for a digital element.

Definition 3.2 (Source)**.**
A ***source****, denoted ω, is composed by both the set of information items present in it and its support. For an information item i, $i \in \omega$ denotes that i is present in source ω.*
The theoretical ***set of sources*** *is noted Ω.*

Considering digital context which is central to our work, a source can be of many kind, from a textual unstructured file, to a database or a worldwide used knowledge graph. There can also be various supports, even for the same type of data source. For instance, the support of a file could be a PDF reader, or a page of a website if the file is online. Similarly, the support of a knowledge graph could be a SPARQL endpoint and all the infrastructure allowing to access it (including possible restriction of access, authentification...), but it can also just be a RDF dump to download.

3.1.3 Accessibility

One of the main component of transparency is accessibility. We first propose a minimalist definition.

Definition 3.3 (Accessibility of a source)**.**
A source ω is ***accessible****, noted* accessible(ω)*, iff there exists an "open, free and universally implementable" protocol [Wil+16] through which ω is retrievable.*

This definition is mainly focused on the support of source ω without any consideration on its content. Furthermore, we choose to use the notion of access "protocol" which is "digital source" oriented. The advantage is to clearly specify what is expected while leaving the possibility of a transposition to other types of sources. For example, considering a paper document, its location must be known, it must be possible to access and read it without hindrance.

Several definitions of the literature impose other criteria for accessibility of a source as for example the necessity of a "license to use" (see 2.2.2). In a general context, the "license to use" is of major importance since it specifies what can or cannot be done with the data. However, because in our context the focus is on information gathering, and notably to consult and not to share nor to modify it, the need of a license does not appear as fundamental. The proposed definition of accessibility is voluntarily minimalist to highlight the strictly necessary elements. Of course, variants can be introduced to take into account the need for a license, but also for identification, authentification, etc.

Accessing a source is important, but we must mind that the objective is to access the information it contains. Indeed, all definitions of accessibility require that all information are retrievable at any time. We will be less demanding, only requiring that information on which we focus is accessible. For an information item to be effectively accessible, some possible obstacles remain. First, and obviously, the source must agree to open access to the information. But this is still not enough. It is also necessary that this access is reasonably easy (not subject to an authorization to be submitted and validated by several authorities), that this data is in an interpretable and usable format (physical or logical format, e.g. poor quality scan, accessible but hand-written document, excessive anonymization), etc. All these obstacles have in common to make access to information more difficult, or to make use of a more easily quantifiable concept, to make it

slower. We therefore propose to introduce the notion of time in the definition of accessibility to information in a source.

Definition 3.4 (Information accessibility in a source)**.**
Let us consider a source ω *and an information item* i. accessible(ω, i) *denotes that* i *is* ***accessible in*** ω *and is defined by:*
Boolean version

$$\text{accessible}(\omega, i) =_{Def} \text{accessible}(\omega) \textit{ and } i \in \omega \textit{ and } \text{time}(\omega.\text{consult}(i)) < \theta$$

Numeric version

$$\text{accessible}(\omega, i) =_{Def} \begin{cases} \dfrac{\theta - \text{time}(\omega.\text{consult}(i))}{\theta} & \textit{if } \text{accessible}(\omega) \textit{ and } i \in \omega \textit{ and } \text{time}(\omega.\text{consult}(i)) < \theta \\ 0 & \textit{else} \end{cases}$$

Where:

- time$(\omega.\text{consult}(i))$ *denotes the necessary time to access* i *in* ω*, and*
- θ *is a reasonable amount of time.*

To consider digital tools, although often cited as solutions to facilitate accessibility, they are not always free of problems related to the difficulty of access to information. Even forgetting the questions of access rights and format, time being an important element in computing, it is fairly well taken into account, but not always to the advantage of the user. For example, when a request lasts too long it can reach its "timeout" and is then abandoned. This avoids too much consumption of resources, but from the user's point of view, it also means that it is not possible to obtain some answers. Thus, some SPARQL endpoints will not respond to certain queries, as for Wikidata for instance [Mal+18]. So, although a server theoretically gives access to the answer to a query, it will not provide it.

The first definition we proposed for accessibility deals with the accessibility of a source. The second one defines the accessibility of an information in a source. The next and last one completes the picture by focusing on the accessibility of an information. Intuitively, an information is accessible if there is at least one source on the planet where it is accessible.

Definition 3.5 (Information accessibility)**.**
Let us consider an information item i. accessible(i) *denotes that* i *is* ***accessible*** *and is defined by:*
Boolean version

$$\text{accessible}(i) =_{Def} \exists \omega \in \Omega \mid \text{accessible}(\omega, i)$$

Numeric version

$$\text{accessible}(i) =_{Def} \max_{\omega \in \Omega}(\text{accessible}(\omega, i))$$

3.2 Transparency based on structured needs only

In this section, without pretending to be exhaustive, we first present three different ways of structuring observation needs. For each case, a corresponding definition of transparency is given, without considering the sources. Then, in the next section, we bring the sources into the picture.

3.2.1 Observation need structured as a set.

The simplest structure which can be used to specify an *observation need* is a set which is assumed to be sound and complete with the observer's requirement. In other terms, an element belongs to the set if and only if it is of interest with respect to the *observation need*, without any consideration on how this set is obtained.

Definition 3.6 (Observation need structured as a set)**.**
The ***observation need****, or simply the* ***o-need****, noted $\mathcal{N}_s$, is a sound and complete set of information w.r.t. the need of the observer.*

Since at this stage there is no explanation on how to build such set, it can be considered as some kind of oracle.

Example 3.1 (Observation need as a set)**.**
Zoe wants more transparency concerning her family. Indeed, she wants to know who are her grand-parents. Contrary to Zoe, the oracle function do know them and define which information she should find. Then, the observation need $\mathcal{N}_s$ associated with her need of transparency is defined as follows. The information items in the o-need are : i_A = "the mother of Zoe's mother is Alice", i_B = "the father of Zoe's mother is Bob", i_C = "the mother of Zoe's father is Cathy", i_D = "the father of Zoe's father is David". So, $\mathcal{N}_s = \{i_A, i_B, i_C, i_D\}$.

As far as the problem of transparency is mainly considered as a question of completeness, considering the need as a set leads to a quite simple definition.

Definition 3.7 (Transparency based on a set)**.**
Let $\mathcal{N}_s$ be an observation need as a set. Transparency w.r.t. this need is defined as follows:

$$\text{transp}_\text{s}(\mathcal{N}_s) =_{Def} \circledast_{i \in \mathcal{N}_s} \text{accessible}(i)$$

where $\circledast$ is an aggregative operator to be specified.

Either Boolean or numeric operators can be used. Assuming accessible(i) returns a number, a simple min function can do some job, so does avg (average), etc. The use of a weighted average (allowing to highlight differences in the importance of some elements compared to others) is also possible and very tempting. This latter requires to enrich the requirement model with a weighting function defined on all elements belonging to $\mathcal{N}_s$. This approach could be developed, but the following proposal comes very close in another way.

3.2.2 Observation need structured by essential items.

One can consider that, to be satisfied, a need has to obtain some major information. Let us qualify such information as *essential*. When an *essential* information item i is obtained, the frontier of known information is pushed back. This can highlight further needs. Indeed, in real world, the need of more information can be issued from a reasoning involving i (which is now part of knowledge) or simply from a questioning about i (e.g. provenance), etc. Whatever, this new information can be considered as essential w.r.t. i such that, this part of the need can be considered as contextual to i and is noted $\mathcal{N}_e^i$. Thus, from essential information to essential information, from context to context, the need appears more structured and a possible construction method begins to take shape. For its part, the initial need $\mathcal{N}_e$, the starting point, is not contextualized to any particular information. From a logical point of view, an absence of

contextualization is equivalent to a contextualization on $TRUE$, the truth. Following this new notation, the initial requirement can be written either $\mathcal{N}_e$ or $\mathcal{N}_e^{TRUE}$.

An observation need $\mathcal{N}_e$ is structured by essentiel items iff there exists an oracle function essential which defines the items i' which are of importance in the context of any item i, noted $\mathcal{N}_e^i$.

Definition 3.8 (Observation need structured by essential items)**.**
Let i be an information item. An ***observation need structured on essential information*** *is defined by an oracle function* essential *such that:*

$$\begin{aligned}\mathcal{N}_e^i.\,\text{essential} : I &\longrightarrow BOOLEAN\\ i' &\longmapsto \big(i' \textit{ is of major interest, essential, in the context of } i\big)\end{aligned}$$

Notations: $\mathcal{N}_e$ is an abbreviation of $\mathcal{N}_e^{TRUE}$; $i' \in \mathcal{N}_e^i$ denotes that in the context of i, i' is of some interest.
The following properties have to be verified:

- $i' \in \mathcal{N}_e^i$ *iff* $(\mathcal{N}_e^i.\,\text{essential}(i')$ *or* $\exists i'' : (i'' \in \mathcal{N}_e^i$ *and* $\mathcal{N}_e^{i''}.\,\text{essential}(i')))$
  From a set point of view, $\mathcal{N}_e^i$ is the closing on essential information.
- *if* $i' \in \mathcal{N}_e^i$ *then* $i \notin \mathcal{N}_e^{i'}$
  No cycle.

Essential information for $\mathcal{N}_e$ therefore corresponds with items of major interest w.r.t. this need (without considering more specific context).

If one is interested in articulating this definition with logical expressions (often used to represent information and knowledge), the following properties can be added, α and β being logical expressions:

- if $\mathcal{N}_e^\alpha.\,\text{essential}(i)$ then $\mathcal{N}_e^{\alpha\vee\beta}.\,\text{essential}(i)$
  If an element is essential for α, then it is essential for any formula whose truth value depends on α.
- $\mathcal{N}_e^\alpha.\,\text{essential}(i)$ iff $\mathcal{N}_e^{\neg\alpha}.\,\text{essential}(i)$
  The essentiality of an element i is not related to the truth value of α (is α true or false), even if i may be of major interest to determine if α is true or not.

Example 3.2 (Need based on the notion "essential")**.**
Let i = "The average price per square meter for an apartment rental in Toulouse is 10 euros per month" be an information item considered as essential for $\mathcal{N}_e$, as well as i_1 = "i was produced in a study on apartments in Toulouse in 2021". Some examples of essential information in the context of i can be: i_2 = "John Sulivan is the author of i", i_3 = "dataSet457 was used for i production". i_1 is also an essential information item in the context of i as it indicates where i could be found in the first place. Some of these last may also have essential information in their context. For example, for i_3 it may be essential to know that i_4 = "Organization Twelve was in charge of dataSet475 collect", as well as i_5 = "dataSet457 was collected in 2020".

Stopping the example here, we have:

- $\mathcal{N}_e.\,\text{essential}(i)$, $\mathcal{N}_e.\,\text{essential}(i_1)$, $\mathcal{N}_e^i.\,\text{essential}(i_1)$, $\mathcal{N}_e^i.\,\text{essential}(i_2)$, $\mathcal{N}_e^i.\,\text{essential}(i_3)$, $\mathcal{N}_e^{i_3}.\,\text{essential}(i_4)$, $\mathcal{N}_e^{i_3}.\,\text{essential}(i_5)$, *and*
- $\{i_4, i_5\} \subseteq \mathcal{N}_e^{i_3}$, $\{i_1, i_2, i_3, i_4, i_5\} \subseteq \mathcal{N}_e^i$ *and* $\{i, i_1, i_2, i_3, i_4, i_5\} \subseteq \mathcal{N}_e$.

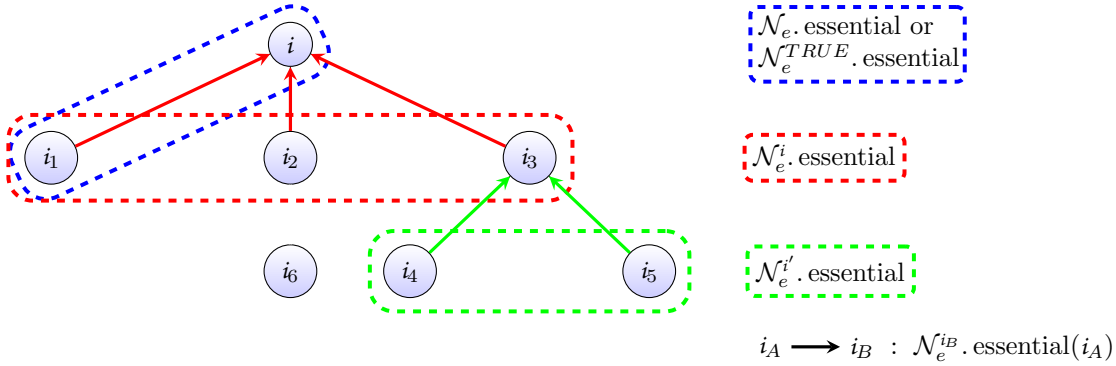


Figure 2: Illustration of an observation need structured w.r.t. essential information

Of course, the observer is aware of the existence of other information, as for example i_6 = "the computer used by John Sulivan to produce i is equipped with an intel core i5 processor", but in our example, the observer does not consider it as essential.

Compared to an observation need defined as a set, here the calculation of transparency can take advantage of how the need is structured.

Definition 3.9 (Transparency based on essential information)**.**
Let $\mathcal{N}_e^i$ be an observation need structured by essential items. Transparency w.r.t. $\mathcal{N}_e^i$ is defined as follows:

$$\text{transp}_\text{e}(\mathcal{N}_e^i) =_{Def} \text{accessible}(i) \circledast \left(\underset{i' | \mathcal{N}_e^i . \text{ essential}(i')}{\circledast} \left(\text{transp}_\text{e}(\mathcal{N}_e^{i'}) \right) \right)$$

Where $\circledast$ is an aggregative operator to be specified.
Reminder: $\mathcal{N}_e = \mathcal{N}_e^{TRUE}$. Furthermore, there is no accessibility question about $TRUE$, so

$$\text{transp}_\text{e}(\mathcal{N}_e) =_{Def} \underset{i' | \mathcal{N}_e . \text{ essential}(i')}{\circledast} \left(\text{transp}_\text{e}(\mathcal{N}_e^{i'}) \right)$$

Clearly, compared to the simple set approach, taking into account the way the need structures the elements improves accuracy.

However, two questions remain open. The first one is again the choice of an aggregation function. As for a simple set, the choice is vast min, max, avg... depending on how restrictive one wants to be. But a second question, possibly a prerequisite to the choice of an aggregation function, arises: "how to define what is essential when focusing on a piece of information?".

3.2.3 Observation need structured by questions and answers.

The question raised at the end of the previous sub-section is in fact a variant of the questions that arise in many contexts (journalism, police investigations, troubleshooting analysis, problem solving): "How to proceed to collect basic information, without forgetting any?". This kind of information gathering problem is extremely common and methods have been proposed to help solve it. The five Ws or 5W [Wik] (and its variations) is one of them. Its principle is quite simple: it lists basic questions (*Who*, *What*, *When*, *Where*, *Why* but also sometimes *How*...) that should

have factual answers (the answer can't be a simple "Yes" or "No"). In this subsection, we draw on this approach to structure the requirement using the concepts of questions and answers.

The approach explored here is inspired from 5W method. The notions of questions and answers are introduced to structure a need:

Definition 3.10 (Observation need structured by questions and answers)**.**
Let i be an information item. Let Q be the set of all possible questions (possibly a question in natural language). And let $q \in Q$ be a question.

*An **observation need**, focused on i structured by questions and answers, noted $\mathcal{N}^i_{qa}$, is defined by two oracle functions* required *and* relevant *such that:*

- required *determines which question is essential w.r.t. $\mathcal{N}^i_{qa}$, i.e. the question for which an answer is required.*

$$\begin{aligned} \mathcal{N}^i_{qa}.\,\text{required} : Q &\longrightarrow BOOLEAN \\ q &\longmapsto \big(\textit{answers to } q \textit{ are required by } \mathcal{N}^i_{qa}\big) \end{aligned}$$

- relevant *recognizes relevant answer to a question w.r.t. $\mathcal{N}^i_{qa}$.*
  Note that a question can have several answers.

$$\begin{aligned} \mathcal{N}^i_{qa}.\,\text{relevant} : Q \times 2^I &\longrightarrow BOOLEAN \\ (q, A) &\longmapsto \big(\textit{the answer } A \textit{ to } q \textit{ provides relevant input w.r.t. } \mathcal{N}^i_{qa}, \\ &\qquad \textit{and } A \textit{ is minimal.}\big) \end{aligned}$$

Notations: $\mathcal{N}_{qa}$ is an abbreviation of $\mathcal{N}^{TRUE}_{qa}$; $i' \in \mathcal{N}^i_{qa}$ denotes that in the context of i, i' is of some interest (regardless how it has been obtained - which answer to which question related to which item).
The following properties have to be verified:

- $i' \in \mathcal{N}^i_{qa}$ iff $\quad (\exists q \in Q, \exists A \in 2^I \mid \mathcal{N}^i_{qa}.\,\text{required}(q) \text{ and } \mathcal{N}^i_{qa}.\,\text{relevant}(q, A) \text{ and } i' \in A)$
  or
  $(\exists i'' \in \mathcal{N}^i_{qa} : i' \in \mathcal{N}^{i''}_{qa})$
  From a set point of view, $\mathcal{N}^i_{qa}$ is the closing on relevant answers to all suitable questions.

- if $i' \in \mathcal{N}^i_{qa}$ then $i \notin \mathcal{N}^{i'}_{qa}$
  No cycle.

If one is interested in articulation with logical expressions (often used to represent information and knowledge), the following properties can be added. Let α, β be logical expressions.

- if $\mathcal{N}^\alpha_{qa}.\,\text{required}(q)$ then $\mathcal{N}^{\alpha \vee \beta}_{qa}.\,\text{required}(q)$
  If a question is interesting w.r.t. α, then it must remain interesting for any formula whose truth value depends on α.

- $\mathcal{N}^\alpha_{qa}.\,\text{required}(q)$ iff $\mathcal{N}^{\neg\alpha}_{qa}.\,\text{required}(q)$
  The fact that a question is interesting w.r.t. α does not depend on the truth value of α, even if the objective of the question is to determine if α is true or not.

To enlight the links with the approach based on essential information, we can say:

$$\{i' \mid \mathcal{N}^i_e.\,\text{essential}(i')\} = \bigcup_{\substack{q,A \mid \mathcal{N}^i_{qa}.\,\text{required}(q) \\ \mathcal{N}^i_{qa}.\,\text{relevant}(q,A)}} (A)$$

Example 3.3 (Need based on the notion "essential")**.**
Let q = "What is the average price per month per square meter for an apartment rental in Toulouse?" be a required question for $\mathcal{N}_e$. Obviously, i = "The average price per square meter for an apartment rental in Toulouse is 10 euros per month" is relevant to answer q.

Some examples of required question in the context of i can be: q_1 = "Who is the author of i?", q_2 = "What was used to produce i?". Relevant answers to these questions might be i_1 = "John Sulivan is the author of i" for q_1 and for q_2: i_2 = "dataSet457 was used for i production" and i_3 = "method AZT was used for i production". i_3 and i_4 are two possible answer to q_2.

Some of these last may also have required questions in their context. For example, for i_2 it may be required to answer the question q_3 = "Who is in charge of dataSet457?" or q_4 = "When was dataSet457 collected?". With the following relevant answers i_4 = "Organization Twelve was in charge of dataSet475 collect" for q_3, and for q_4: it is possible to deduce that dataSet475 was collected in 2020 thanks to both i_4 and i_5 = "Organization Twelve existed from February 2020 to October 2020". i_4 and i_5 are only relevant to q_4 if considered together.

Stopping the example here, we have:

- $\mathcal{N}_{qa}.\text{required}(q)$, $\mathcal{N}_{qa}.\text{relevant}(q, \{i\})$
  $\mathcal{N}^{i}_{qa}.\text{required}(q_1)$, $\mathcal{N}^{i}_{qa}.\text{relevant}(q_1, \{i_1\})$
  $\mathcal{N}^{i}_{qa}.\text{required}(q_2)$, $\mathcal{N}^{i}_{qa}.\text{relevant}(q_2, \{i_2\})$ *and* $\mathcal{N}^{i}_{qa}.\text{relevant}(q_2, \{i_3\})$
  $\mathcal{N}^{i_2}_{qa}.\text{required}(q_3)$, $\mathcal{N}^{i_2}_{qa}.\text{relevant}(q_3, \{i_4\})$
  $\mathcal{N}^{i_2}_{qa}.\text{required}(q_4)$, $\mathcal{N}^{i_2}_{qa}.\text{relevant}(q_4, \{i_4, i_5\})$.
- $\{i_4, i_5\} \subseteq \mathcal{N}^{i_2}_{qa}$, $\{i_1, i_2, i_3, i_4, i_5\} \subseteq \mathcal{N}^{i}_{qa}$ *and* $\{i, i_1, i_2, i_3, i_4, i_5\} \subseteq \mathcal{N}_{qa}$.

Before going further, let us see how the aggregation method in the definition of transparency can be refined to take the most of the new structure.

Definition 3.11 (Transparency based on questions and answers)**.**
Let $\mathcal{N}^{i}_{qa}$ be an observation need structured by questions and answers. Transparency is defined as a succession of aggregation functions that follow the structure of the requirement into questions and answers. Hence, transparency w.r.t. this need is defined as follows:

$$\text{transp}_{\text{qa}}(\mathcal{N}^{i}_{qa}) =_{Def} \text{accessible}(i) \circledast \left(\underset{q|\mathcal{N}^{i}_{qa}.\,\text{required}(q)}{\textcircled{Q}} \left(\underset{A|\mathcal{N}^{i}_{qa}.\,\text{relevant}(q,A)}{\textcircled{A_q}} \left(\underset{i' \in A}{\textcircled{I}} (\text{transp}_{\text{qa}}(\mathcal{N}^{i'}_{qa})) \right) \right) \right)$$

Where

- $\circledast$ *is an aggregative operator to be specified.*
- $\textcircled{Q}$ *aggregates the quality of transparency obtained for all the questions asked by the observer.*
- $\textcircled{A_q}$ *aggregates the quality of transparency obtained for all the answers obtained for one question.*
- $\textcircled{I}$ *aggregates the quality of transparency obtained for information being part of one answer.*

In the previous definition, $\textcircled{A_q}$ is build considering that a question can have several answers. For example, when asking about funding for a project, there may be several sources of funding; when asking how a result can be achieved, there may be several methods to achieve the same

result, etc. In the first given example, the different sources of fundings have to be considered as a conjunctive list. On the contrary, when several methods are proposed to obtain the same result, it is more natural to consider them as a disjunctive list. So, clearly, the choice of the aggregation method depends on the query, which explain why $\textcircled{A_q}$ is indiced.

Assuming that all the questions are built to obtain answers to be considered as a disjunctive list, a possible instantiation could be:

$$\underset{q \in \mathcal{N}^i_{qa}.\,\mathrm{required}(q)}{\mathrm{avg}} \left(\underset{A \;|\; \mathcal{N}^i_{qa}.\,\mathrm{relevant}(q,I')}{\max} \left(\min_{i' \in A} \left(\mathrm{transp}_{\mathrm{qa}}(\mathcal{N}^{i'}_{qa}) \right) \right) \right)$$

Assuming that all the questions are built to obtain answers to be considered as a conjunctive list, a possible instantiation could be:

$$\underset{q \in \mathcal{N}^i_{qa}.\,\mathrm{required}(q)}{\mathrm{avg}} \left(\underset{A \;|\; \mathcal{N}^i_{qa}\,\mathrm{relevant}(q,I')}{\min} \left(\min_{i' \in A} \left(\mathrm{transp}_{\mathrm{qa}}(\mathcal{N}^{i'}_{qa}) \right) \right) \right)$$

or simply

$$\underset{q \in \mathcal{N}^i_{qa}.\,\mathrm{required}(q)}{\mathrm{avg}} \left(\underset{A | \mathcal{N}^i_{qa} \;|\; \mathrm{relevant}(q,I')}{\min_{i' \in A}} \left(\mathrm{transp}_{\mathrm{qa}}(\mathcal{N}^{i'}_{qa}) \right) \right)$$

Compared to the previous approach based on the notion of essential information, the need is clearly much more structured and the aggregation much more precise, but this is not the only point. Indeed, proceeding by questions and answers allows to get closer to the natural behavior of a user seeking to collect data. Moreover, from a technical point of view, the most direct possible implementation of these definitions is the implementation of the oracle functions. If $\mathcal{N}_e.\,\mathrm{essential}$ is clearly very difficult to implement, the identification of the questions to be asked to obtain information ($\mathcal{N}_{qa}.\,\mathrm{required}$) seems within reach. For example, it seems possible to build generic questions to enquire on certain classes of information as author, dates, method, etc. Moreover, knowledge, information or data management tools come with queries facilities, which are by nature very close to the notion of question. Finally, analyzing the obtained answers can be difficult, depending on the desired level of precision, but it seems possible to have some achievements on this side too. In conclusion, this structure is stronger than previous ones, closer to a natural behavior and also seems to facilitate numeric implementation.

3.3 Transparency and sources.

Last but not least, we have to consider sources. Indeed, one of our objectives is to evaluate the transparency of a source.

While information can be obtained from a variety of sources, it is not reasonable to look for all information of interest in all sources. It is even less reasonable to judge the transparency of a source by comparing the information it provides against all complete need. Indeed, not all sources provide the same information, not all sources have the same objectives. We must therefore differentiate between what we can expect from one source and what we can expect from another.

This leads to introduce a new oracle function:

Definition 3.12 (Expected information from a source)**.**
Let us consider a source $\omega \in \Omega$ and an information item $i \in \mathcal{I}$. Let $\mathcal{N}$ be a need.
$\mathcal{N}.\,\mathrm{expected}(\omega, i)$ *denotes that i is* ***expected to be accessible*** *in ω.*

This function expected is defined no matter the structure of the need chosen. Furthermore, whether or not a source is expected to provide information is completely independent of any context, of any focused information. Thus, even while focusing on i, $\mathcal{N}.\,\text{expected}$ will be used to determine if a source is expected to make some information accessible. In other terms, even if $\mathcal{N}^i.\,\text{expected}$ can be used, it is equivalent to $\mathcal{N}.\,\text{expected}$ use of which is recommended.

Clearly, determining whether a source should provide information requires strong knowledge. Here, we will assume that it is $\mathcal{N}$, the need itself, that provides the function expected. Therefore, formally, we consider it as an oracle function.

Mixing the source with the previous proposal leads to:

Definition 3.13 (Transparency of a source based on a set)**.**
Let $\mathcal{N}_s$ be a observation need as a set, and ω be a source.
Transparency of the source w.r.t. this need is defined as follows:

$$\text{transp}_\text{s}(\mathcal{N}_s, \omega) =_{Def} \bigcircledast_{\substack{i \in \mathcal{N}_s \\ \mathcal{N}_s.\,\text{expected}(\omega, i)}} \text{accessible}(\omega, i)$$

Definition 3.14 ([Transparency of a source based on essential information)**.**
Let $\mathcal{N}_e^i$ be a structured observation need, and ω be a source.
Transparency of the source w.r.t. this need is defined as follows:

$$\text{transp}_\text{e}(\mathcal{N}_e^i, \omega) =_{Def} \text{accessible}(\omega, i) \circledast \left(\bigcircledast_{\substack{i' | \mathcal{N}_e^i.\,\text{essential}(i') \\ \mathcal{N}_e.\,\text{expected}(\omega, i')}} \left(\text{transp}(\mathcal{N}_e^{i'}, \omega) \right) \right)$$

Definition 3.15 (Transparency of a source based on questions and answers)**.**
Let $\mathcal{N}_{qa}^i$ be a observation need structured by questions and answers, and ω be a source.
Transparency of the source w.r.t. this need is defined as follows:

$$\text{transp}_\text{qa}(\mathcal{N}_{qa}^i, \omega) =_{Def} \text{accessible}(\omega, i) \circledast \left(\bigcircledast_{\substack{q \in Q \\ \mathcal{N}_{qa}^i.\,\text{required}(q)}} \left(\bigcircledast_{\substack{A \in 2^I \\ \mathcal{N}_{qa}^i.\,\text{relevant}(q, A)}} \left(\bigcircledast_{\substack{i' \in A \\ \mathcal{N}_{qa}.\,\text{expected}(\omega, i')}} (\text{transp}(\mathcal{N}_{qa}^{i'}, \omega)) \right) \right) \right)$$

Reminder: for the last two definitions, $\mathcal{N} = \mathcal{N}^{TRUE}$.

3.4 Discussion

As indicated in the synthesis of the section 2, transparency is based on openness, accessibility and completeness. The different definitions presented in this section propose different solutions for the verification of completeness, which requires the definition of the need. Of course, they all rely on the accessibility of information. This one is defined by the fact that information is actually present in the source (openness), that the source is accessible and the information retrievable in reasonable time (accessibility). Openness of the source is not evaluated because we prefer to focus only on the information on interest just like we do with the accessibility.

As in our context the notion of source is unavoidable, for the choice of the definition of transparency to use the set of candidates is limited to the last three definitions.

First of all, these last differ in the way the need is defined: a “set”, a “set structured by the notion of essential information”, or a “set structured by questions and answers”. From a practical

and implementation point of view, the need will have to be specified. Although formally suitable, the notion of "set" leaves the question "How to obtain the information required by the need?" totally open. Although it improves the structuration of the need, the notion of essentiality only slightly shifts the question to "How to get the essential information required by the need?". In comparison, the third definition provides an answer: "The information is obtained by asking questions." but leads to another question: "Which questions to ask?". It turns out that this last question brings us back to the well-known problem of information gathering. In practice, different tools have been developed to help solve this problem. One of the best known is the 5W method (and its variants) [Wik] which helps to determine the important questions to ask w.r.t. a need. In addition, in the literature some works [OH20] are also closely related to this method. This last definition is therefore the one that relates the best to existing works. Moreover, in our context of knowledge graphs, the proximity of the notion of "question" to that of "query" sounds like a promise of an intuitive implementation.

According to these considerations, definition 3.15 is the one that brings the most precision in the definition of transparency, with a definition of the need closest to existing work and with the best perspectives of implementation. Our efforts will therefore focus on the avenues it opens up.

4 Formal definition of transparency of knowledge graphs

In this section, the previous definitions are adapted for a knowledge graph, we only focus on the ones implemented as RDF graph.

4.1 Preliminary notions

Let us start by introducing some definitions and notations. For more details, see [PAG09] and [HSP13].

Definition 4.1 (RDF Triple)**.**
Let $\mathcal{I}$ be the set of IRIs, $\mathcal{B}$ the set of blank nodes, and $\mathcal{L}$ the set of literals.
An RDF triple is a tuple $(s, p, o) \in (\mathcal{I} \cup \mathcal{B}) \times \mathcal{I} \times (\mathcal{I} \cup \mathcal{B} \cup \mathcal{L})$, where s is the subject, p the predicate and o the object [PAG09].

Definition 4.2 (RDF graph)**.**
An RDF graph, or RDF dataset, is a a set of RDF triples [PAG09].

For the following, we will use the term knowledge graph, denoted ω, and we will consider it with both the RDF graph and its support. It is possible to query a knowledge graph with SPARQL which is a graph-matching query language. The core part of a SPARQL query are graph patterns. The objective is to match the graph patterns against the knowledge graph in order to solve the query. Hence, a query is based on triple patterns: it provides one or more patterns where one or more elements of the triples are variables. A query is denoted q and $\mathcal{Q}$ is the set of queries.

Definition 4.3 (Triple Pattern)**.**
Let $\mathcal{V}$ be the set of (query) variables. A triple pattern is a tuple $(s, p, o) \in (\mathcal{V} \cup \mathcal{I} \cup \mathcal{B} \cup \mathcal{L}) \times (\mathcal{V} \cup \mathcal{I}) \times (\mathcal{V} \cup \mathcal{I} \cup \mathcal{B} \cup \mathcal{L})$[2] [HSP13].
Let q be a query, then $Triples(q)$ denotes the set of all triples patterns involved in the query q. And, $var(q)$ is the set of query variables occurring in q.

[2]Notice that s could theoretically be a literal according to the W3C specification of SPARQL, even if it will not match any RDF triple.

Definition 4.4 (Solution Mapping)**.**
A solution mapping μ is a partial function from a set of variables to a set of RDF terms, i.e. $\mu : \mathcal{V} \to \mathcal{I} \cup \mathcal{B} \cup \mathcal{L}$ [HSP13; PAG09].
By a common abuse of notation, $\mu(t)$ denotes the RDF triple obtained by replacing the variables in a triple pattern t according to μ By extension, $\mu(Triples(q)) = \{\mu(t) | t \in Triples(q)\}$ is the set of all triples obtained by replacing every variable in every triple pattern in q according to μ.

Then given a knowledge graph, let us define a solution of a query. The solution should not be confused with the result of the query which is the output, defined after. The solution is the mapping that, after blank nodes have been replaced by IRIs, literals or blank nodes, makes the triple patterns match with the RDF triples of the graph. The following definition is inspired from the definition of Basic Graph Pattern Matching of [HSP13].

Definition 4.5 (Solution)**.**
Let ω be a knowledge graph and q be a query. The solution mapping $\mu : var(q) \to \mathcal{I} \cup \mathcal{B} \cup \mathcal{L}$ is a solution for q from ω if it exists σ an RDF instance mapping (used to replace blank nodes into IRIs, literals or blank nodes), such as $\mu(\sigma(Triples(q)))$ is a subgraph of ω. Then the set of solutions for q from ω is $Sol_\omega(q) = \{\mu$ a solution mapping $| \exists \sigma, \mu(\sigma(Triples(q)) \subseteq \omega)\}$.

Definition 4.6 (Result)**.**
*The **result** of a query q on a knowledge graph ω is the output of the query executed on ω. It is denoted* result(ω, q)*. It is either a Boolean for ASK queries, or a new RDF graph for DESCRIBE and CONSTRUCT queries, or a list of bindings, i.e. values matching variables for SELECT queries.*

Property 4.1 (Non-empty result)**.**
*A result is not empty (*result$(\omega, q) \neq \varnothing$*) when:*

- *for ASK queries: it is always not empty ;*
- *for DESCRIBE/CONSTRUCT queries: it contains a non empty graph ;*
- *for SELECT queries: it contains at least one binding.*

4.2 Formalization of transparency adapted to knowledge graphs

The notion of information item was central in the definition of an observation need, it can be adapted for knowledge graphs, where queries and triples can both be information.

Definition 4.7 (Information item)**.**
An information item is either a set of triples of ω, or the result of a query.

Then, we can define the *accessibility* of a knowledge graph. The accessibility of the source, accessible(ω), does not change from Definition 3.3, but the accessibility of an information item now applies to a query, and the notion of execution time appears naturally. We also require to have a non-empty result to the query to consider it accessible, for one generally asks a query in order to get a result.

Definition 4.8 (Accessibility of a query in a knowledge graph)**.**
Let us consider a knowledge graph ω and a query q. accessible(ω, q) *denotes that q is **accessible** in ω, and is defined by:*

$$\text{accessible}(\omega, q) =_{Def} \text{accessible}(\omega) \text{ and } \text{result}(\omega, q) \neq \varnothing \text{ and } \text{time}(\omega.\text{execute}(q)) \leqslant \theta$$

Where θ is a reasonable amount of time. time$(\omega.\text{execute}(q))$ *is the execution time of the query, i.e. the time to get all results of the query.*

Typically, θ can be the execution timeout of the associated SPARQL endpoint.

Now, it is time to define the observation need in the context of knowledge graphs. Previously, we defined it globally or with regard to an information item and structured it according to information items also. In order to define the information essential for a need, we either used an oracle function essential which directly gives a set of information, or a function required which returns a set of essential question that lead to essential information items.

We adapt the notion of need to knowledge graphs, queries arising naturally to replace questions. It is also more natural to evaluate accessibility regarding queries. Queries involve triples to be solved while questions involve information items. Then, it is possible to define a global need $\mathcal{N}_q$ thanks to a set of queries, and needs local to a triple $\mathcal{N}_q^t$ thanks to queries about the triple.

Definition 4.9 (Need structured by queries)**.**
*An **observation need**, focused on t, **structured by queries** , $\mathcal{N}_q^t$, is defined by two functions,* required *and* expected*. First,* required *asserts if the result of a query is required or not w.r.t. the need.*

$$\begin{aligned} \mathcal{N}_q^t.\,\text{required} : \mathcal{Q} &\longrightarrow BOOLEAN \\ q &\longmapsto \left(\textit{any result of } q \textit{ is required by } \mathcal{N}_q^t\right) \end{aligned}$$

Then, expected denotes that the query should get a result in the knowledge graph. We assume that it means that the query is accessible according to the definition 4.8.

$$\begin{aligned} \mathcal{N}_q.\,\text{expected} : \Omega \times \mathcal{Q} &\longrightarrow BOOLEAN \\ (\omega, q) &\longmapsto \left(q \textit{ is expected to be accessible in } \omega \textit{ according to } \mathcal{N}_q\right) \end{aligned}$$

In order to structure the need, the global need $\mathcal{N}_q$ (or $\mathcal{N}_q^{TRUE}$) is obtained without any focus on a specific triple. And then, local needs $\mathcal{N}_q^t$ are defined, i.e. needs specific to triples t of the knowledge graph and allows to define new queries after considering the results obtained so far.

When defining a need specific to a triple, the required queries may focus on the whole triple, for instance "who adds the triple to the knowledge graph?". Or they may also focus on one entity of the triple, for instance ask for additional information about the subject of the triple.

Furthermore, only the function required bears the hierarchical structure of the need. Hence, the function expected is independent of the level on which it is considered: it is not necessary to define it with regard to a specific triple. This means the function is the same for $\mathcal{N}_q$ or $\mathcal{N}_q^t$ for all triples. So in practice, we only define expected once, at the scale of the main need $\mathcal{N}_q$.

In this definition, we do not assume to know the answer nor that one set of triples is the answer of a question. To draw parallels with the definition 3.10 of a need structured by questions, the result of a query is always the relevant answer to it. So, we do not need the function relevant here, the execution of a query will give it to us. This definition does not use oracle functions anymore, but only queries defined by a user.

Example 4.1 (Need structured by queries)**.**
Zoe wants to know who her grand-parents are. She knows her mother and her father and she also knows that they are respectively British and French. Hence, she has access to two different sources, ω_1 = British register office and ω_2 = French register office. As her father always lied to her concerning her grand-parents, she has a strong need of verifiability. Hence, Zoe wants ω_1 and ω_2 to be transparent concerning her family.

Zoe does not know who her grand-parents are. But she can define queries to identify them. Then, her need is defined as follows. For instance, the required queries could be $\{q \in \mathcal{Q} \mid \mathcal{N}_q.\,\text{required}(q)\} = \{\texttt{A}, \texttt{B}, \texttt{C}, \texttt{D}\}$*, where they are written as follows, knowing the vocabulary used by the two sources.*

```
A:SELECT ?grandmotherM                          B:SELECT ?grandfatherM
  WHERE  ?grandmotherM :mother_of ?mother .       WHERE  ?grandfatherM :father_of ?mother .
         ?mother :mother_of :Zoe .                        ?mother :mother_of :Zoe .

C:SELECT ?grandmotherF                          D:SELECT ?grandfatherF
  WHERE  ?grandmotherF :mother_of ?father .       WHERE  ?grandfatherF :father_of ?father .
         ?father :father_of :Zoe .                       ?father :father_of :Zoe .
```

She knows that the information she is looking for are not in the two sources simultaneously and she knows where to find each of them. The ones concerning Zoe's mother are in the British register office (ω_1*) and the ones concerning Zoe's father are in the French register office (*ω_2*). Hence,* $\{q \in \mathcal{Q} \mid \mathcal{N}_q.\,\text{expected}(\omega_1, q)\} = \{\texttt{A}, \texttt{B}\}$ *et* $\{q \in \mathcal{Q} \mid \mathcal{N}_q.\,\text{expected}(\omega_2, q)\} = \{\texttt{C}, \texttt{D}\}$

Then, she wants some evidence about the triples in source ω_2*, hence she is looking for a scan of the family record book of her grand-parents where her father should be mentioned as their child. Zoe knows who her father is, so the triple pattern* t_f`?father :father_of :Zoe` *does not implies other query. But the other triple patterns* `?grandmotherF :mother_of ?father` *and* `?grandfatherF :father_of ?father` *implies new queries for Zoe. One the first queries* `C` *and* `D` *are executed, she finds that* `?grandmotherF = :Cathy`*,* `?grandfatherF = :David` *and* `?father = :zoe_father`*. So she defines queries on the triples* t=`:Cathy :mother_of :zoe_father` *and* t'=`:David :father_of :zoe_father`*.*

Then, the needs specific to these triples are $\{q \in \mathcal{Q} \mid \mathcal{N}_q^t.\,\text{required}(q)\} = \{\texttt{T}\}$ *and* $\{q \in \mathcal{Q} \mid \mathcal{N}_q^{t'}.\,\text{required}(q)\} = \{\texttt{T'}\}$*. Where,* `T` *and* `T'` *are defined as follows.*

```
T: SELECT ?entity1
   WHERE  :Cathy :family_record_book ?entity1 .
T':SELECT ?entity2
   WHERE  :David :family_record_book ?entity2 .
```

To assess the transparency of a source, the need is evaluated against each solution in each query and each triple composing the result of a query. So, we formulate an hypothesis: all triple patterns involved in a query are useful, i.e. the set $Triples(q)$ is minimal.

The definition of transparency differs slightly from before. Indeed, accessibility is not evaluated against an information item, nor the triple t such that $\mathcal{N}_q^t$, but against all queries required by t. In fact, the accessibility of the triple has already been tested before, as part of the solution of a query.

Transparency of a knowledge graph w.r.t. a need (specific to a triple, or in absolute) is defined by the aggregation on all required queries (specific to the triple or in absolute) of the accessibility of the query and the aggregation on all triples involved in the resolution of the query of the transparency of the need related with this triple.

Definition 4.10 (Transparency of a knowledge graph w.r.t. a need)**.**
Let ω *be a knowledge graph, and* $\mathcal{N}_q$ *be a need structured by queries. Transparency w.r.t. this need is defined as follows:*

$$\text{transp}_q(\mathcal{N}_q^t, \omega) =_{Def} \underset{\substack{q|\mathcal{N}_q^t.\,\text{required}(q) \\ \mathcal{N}_q.\,\text{expected}(\omega,q)}}{\textcircled{Q}} \left(\text{accessible}(\omega, q) \circledast \left(\underset{\mu \in Sol(q)}{\textcircled{A}} \left(\underset{t' \in \mu(Triple(q))}{\textcircled{I}} \text{transp}_q(\mathcal{N}_q^{t'}, \omega) \right) \right) \right)$$

It is possible to break this formula down into two parts, one centered on the main need $\mathcal{N}_q$ and this other one centered on a query.

Definition 4.11 (Transparency of a knowledge graph w.r.t. a need and a query)**.**
Let ω be a knowledge graph, transparency of ω with regard to a need $\mathcal{N}_q$ depends on the transparency of the result of queries that are essential according to the need, and expected to be in the source.

$$\text{transp}_q(\mathcal{N}_q, \omega) =_{Def} \underset{\substack{q|\mathcal{N}_q.\,\text{required}(q)\\ \mathcal{N}_q.\,\text{expected}(\omega,q)}}{\text{Ⓠ}} \big(\text{transp}_q(\mathcal{N}_q, \omega, q)\big)$$

Where transparency of the result of a query q in a knowledge graph ω w.r.t. a need $\mathcal{N}_q$ is:

$$\text{transp}_q(\mathcal{N}_q, \omega, q) =_{Def} \text{accessible}(\omega, q) \circledast \left(\underset{\mu \in Sol(q)}{\text{Ⓐ}} \left(\underset{t \in \mu(Triple(q))}{\text{Ⓘ}} \left(\underset{\substack{q'|\mathcal{N}_q^t.\,\text{required}(q')\\ \mathcal{N}_q.\,\text{expected}(\omega,q')}}{\text{Ⓠ}} (\text{transp}_q(\mathcal{N}_q, \omega, q')) \right)\right)\right)$$

Theorem 4.1 (Equivalence)**.**
Definition 4.10 and Definition 4.11 are equivalent.

Proof. Let ω be a knowledge graph, and $\mathcal{N}_q$ be a need based on queries. Transparency in Definition 4.11 is defined as follows:

$$\begin{aligned}
\text{transp}_q(\mathcal{N}_q, \omega) &= \underset{\substack{q|\mathcal{N}_q.\,\text{required}(q)\\ \mathcal{N}_q.\,\text{expected}(\omega,q)}}{\text{Ⓠ}} \big(\text{transp}_q(\mathcal{N}_q, \omega, q)\big)\\
&= \underset{\substack{q|\mathcal{N}_q.\,\text{required}(q)\\ \mathcal{N}_q.\,\text{expected}(\omega,q)}}{\text{Ⓠ}} \Bigg(\text{accessible}(\omega, q) \circledast\\
&\quad \left(\underset{\mu \in Sol(q)}{\text{Ⓐ}} \left(\underset{t \in \mu(Triple(q))}{\text{Ⓘ}} \left(\underset{\substack{q'|\mathcal{N}_q^t.\,\text{required}(q')\\ \mathcal{N}_q.\,\text{expected}(\omega,q')}}{\text{Ⓠ}} (\text{transp}_q(\mathcal{N}_q, \omega, q')) \right)\right)\right)\Bigg)\\
&= \underset{\substack{q|\mathcal{N}_q.\,\text{required}(q)\\ \mathcal{N}_q.\,\text{expected}(\omega,q)}}{\text{Ⓠ}} \left(\text{accessible}(\omega, q) \circledast \left(\underset{\mu \in Sol(q)}{\text{Ⓐ}} \left(\underset{t \in \mu(Triple(q))}{\text{Ⓘ}} \text{transp}_q(\mathcal{N}_q^t, \omega) \right)\right)\right)
\end{aligned}$$

Hence, for the specific case $\mathcal{N}_q$, transparency defined in Definition 4.11 is equivalent to transparency defined in Definition 4.10. Recursively it also works for $\mathcal{N}_q^t$, for all t. So the two definitions are equivalent. □

In practice, it may be difficult to analyse separately every triple in every solution and to define if new queries are required for each one of them. In order to simplify the process, one can choose to define required queries on the triple patterns involved in the queries, before knowing the solutions. The definition of transparency does not change, it will just be the same queries for each solution, but not instantiated with the same triple.

This definition of transparency, and all three others in the previous section, are contextualized considering a particular need. One could also be interested in studying the transparency of a source in the absolute. That does not appear to us to be pertinent. Indeed, we have not find a general definition of transparency which goes beyond the context, so there is no need to look for a formal definition of it. But it is possible to evaluate the transparency of a source in the context of a generic - and explicit - need. It can be created based on classic objectives on provenance, or by analysing frequent information items required in diverse situations for instance. It is also possible to evaluate a general transparency by aggregating its transparency evaluated on smaller needs related to specific contexts. This will be discuss in more details in the section 5.2.

4.3 Example of a need on a classic knowledge graph: Wikidata

In this section, we illustrate the notion of need and the ensuing calculation of transparency on the knowledge graph Wikidata[3].

The main objective is to evaluate transparency of the knowledge graph with regards to itself in a context of accountability: we want to know if it indicates what are its influences, on a financial point of view and on a contribution of content point of view, i.e. who finances it and who contributes to it. Another objective, on which we will not focus, could be the transparency of the knowledge graph concerning its content, its functioning, its usage conditions, through the search of a license, a SPARQL endpoint, the list of vocabularies used, etc.

Considering this global objective, we ask the following questions. And for each of these questions, we will define a query q.

- Who created the knowledge graph? $\leftrightarrow q_{creator}$
- Who finances the knowledge graph? $\leftrightarrow q_{funder}$
- Who has modified the knowledge graph? $\leftrightarrow q_{contributor}$
- Who owns the knowledge graph? $\leftrightarrow q_{owner}$

Then, we have complementary questions: What is the status of the entities financing the KG? If they are companies, who owns them? And if they are foundations, who run them? If there are multiple financing entities, in what proportion each of them finances the KG ? For the following, we will only keep the complementary question:

- What is the status of the entities financing the knowledge graph? $\leftrightarrow q_{funder-status}$

Now let us define the queries on Table 1. The entity corresponding to Wikidata is "wd:Q2013".

After defining and executing the queries and especially q_{funder}, we found the entities financing the knowledge graph. The triples involved are: t_1 : `wd:Q2013 wdt:P859 wd:Q16002567`, t_2 : `wd:Q2013 wdt:P859 wd:Q95`, t_3 : `wd:Q2013 wdt:P859 wd:Q3070474`.

On each of these triples, we define an interesting query $q_{funder-status} = q_{f-s}(t)$. We use the object (`wd:Q16002567`, `wd:Q95` or `wd:Q3070474`) of the focused triple to replace `<funder>` in this query. It is presented as the last query of the Table 1.

Then we can define the need $\mathcal{N}_q$. As only Wikidata is studied, all required queries are also expected. Then, $\mathcal{N}_q$.required is defined as follows:

$$\{q \in \mathcal{Q} \mid \mathcal{N}_q.\text{required}(q)\} = \{q_{creator}, q_{contributor}, q_{owner}, q_{funder}\}$$

[3] `www.wikidata.org`

Table 1: Queries corresponding to the need

$q_{creator}$:

```
SELECT DISTINCT ?prop ?creator
WHERE {
  # Wikidata has creator ?creator:
  { wd:Q2013 wdt:P170 ?creator . }
  UNION {
    # OR similarly for all sub-properties
    #   of creator
    ?obj_prop wdt:P1647 wd:P170 .
    ?obj_prop wikibase:directClaim ?prop .
    wd:Q2013 ?prop ?creator . }
}
```

$q_{contributor}$:

```
SELECT DISTINCT ?prop ?sub ?contributor
WHERE {
  # For properties contributor and author
  VALUES ?obj_prop { wd:P767 wd:P50}
  # Wikidata has for author or
  #   contributor ?contributor
  {?obj_prop wikibase:directClaim ?prop .
   wd:Q2013 ?prop ?contributor . }
  UNION {
    # OR similarly for all their
    #   sub-properties
    ?sub_prop wdt:P1647 ?obj_prop .
    ?sub_prop wikibase:directClaim ?sub .
    wd:Q2013 ?sub ?contributor . }
}
```

q_{owner}:

```
SELECT DISTINCT ?prop ?sub ?owner
WHERE {
  # Wikidata is owned by ?owner
  { wd:Q2013 wdt:P127 ?owner . }
  UNION {
    # OR similarly for all sub-properties
    #   of owned by
    ?obj_sub wdt:P1647 wd:P127 .
    ?obj_sub wikibase:directClaim ?sub .
    wd:Q2013 ?sub ?owner . }
}
```

q_{funder}:

```
SELECT DISTINCT ?prop ?funder
WHERE {
  # Wikidata has sponsor ?funder:
  { wd:Q2013 wdt:P859 ?funder . }
  UNION {
    # OR similarly for all sub-properties
    #   of sponsor
    ?obj_prop wdt:P1647 wd:P859 .
    ?obj_prop wikibase:directClaim ?prop .
    wd:Q2013 ?prop ?funder . }
}
```

$q_{funder-status}(t) = q_{f-s}(t)$:

```
SELECT DISTINCT ?funder ?legalForm
WHERE {
  # Let ?funder be the funder obtained previously
  BIND (<funder> AS ?funder)
  # The funder has for legal form ?legalForm
  { ?funder wdt:P1454 ?legalForm . }
  UNION {
    # OR funder is an instance of ?legalForm
    #   which is itself an instance of legalForm
    ?funder wdt:P31 ?legalForm .
    ?legalForm wdt:P31 wd:Q12047392 .
  }
}
```

We can continue defining the need as the scale of triples t_1, t_2, t_3 mentioned before.

$$\{q \in \mathcal{Q} \mid \mathcal{N}_q^{t_1}.\,\text{required}(q)\} = \{q_{f-s}(t_1)\}$$

$$\{q \in \mathcal{Q} \mid \mathcal{N}_q^{t_2}.\,\text{required}(q)\} = \{q_{f-s}(t_2)\}$$

$$\{q \in \mathcal{Q} \mid \mathcal{N}_q^{t_3}.\,\text{required}(q)\} = \{q_{f-s}(t_3)\}$$

For all other triples t, $\mathcal{N}_q^t$ is empty, which means that there is not any required query q such that $\mathcal{N}_q^t.\,\text{required}(q)$ is true.

Having defined the need, we can evaluate the transparency of Wikidata with regards to our need. Let us start by the transparency of Wikidata w.r.t. the queries. As there is only one source, we omit ω and *expected*.

For $q_1 = q_{creator}$, all triples involved in the solutions are associated with an empty need. Hence the second part of the formula is en empty aggregation. So transparency of query $q_{creator}$ on Wikidata considering the need $\mathcal{N}_q$ is reduced to the accessibility of (the result of) the query.

$$\begin{aligned}\text{transp}_q(\mathcal{N}_q, q_1) &= \text{accessible}(q_1) \circledast \left(\underset{\mu \in Sol(q_1)}{\textcircled{A}} \left(\underset{t \in \mu(Triple(q_1))}{\textcircled{I}} \left(\underset{q' \mid \mathcal{N}_q^t.\,\text{required}(q')}{\textcircled{Q}} (\text{transp}_q(\mathcal{N}_q, q')) \right) \right) \right) \\ &= \text{accessible}(q_1)\end{aligned}$$

For $q_2 = q_{contributor}$, all triples involved in the solutions are also associated with an empty need, hence its transparency on Wikidata considering the need $\mathcal{N}_q$ is also reduced to its accessibility.

$$\text{transp}_q(\mathcal{N}_q, q_2) = \text{accessible}(q_2)$$

$q_3 = q_{owner}$ also follows the same pattern:

$$\text{transp}_q(\mathcal{N}_q, q_3) = \text{accessible}(q_3)$$

However, $q_4 = q_{funder}$ involves three triples with a non empty associated need. The triples t_1, t_2, t_3 all came from different solutions, respectively μ_A, μ_B and μ_C. And conversely, for each solution, there is always only one triple t for which $\mathcal{N}_q^t$ is not empty. $\mathcal{N}_q^{t_1}, \mathcal{N}_q^{t_2}$ and $\mathcal{N}_q^{t_3}$ only consider one query $q_{f-s}(t)$ as required, and none of the triples involved in this query is associated with a non empty need. Hence, the transparency of $q_{f-s}(t)$ is just its accessibility.

$$
\begin{aligned}
\mathrm{transp}_q(\mathcal{N}_q, q_4) &= \mathrm{accessible}(q_4) \circledast \left(\underset{\mu \in Sol(q_4)}{\textcircled{A}} \left(\underset{t \in \mu(Triple(q_4))}{\textcircled{I}} \left(\underset{q' | \mathcal{N}_q^t . \, \mathrm{required}(q')}{\textcircled{Q}} \mathrm{transp}_q(\mathcal{N}_q, q') \right) \right) \right) \\
&= \mathrm{accessible}(q_4) \circledast \Bigg(\left(\underset{t \in \mu_A(Triple(q_4))}{\textcircled{I}} \left(\underset{q' | \mathcal{N}_q^t . \, \mathrm{required}(q')}{\textcircled{Q}} \mathrm{transp}_q(\mathcal{N}_q, q') \right) \right) \\
&\qquad \textcircled{A} \left(\underset{t \in \mu_B(Triple(q_4))}{\textcircled{I}} \left(\underset{q' | \mathcal{N}_q^t . \, \mathrm{required}(q')}{\textcircled{Q}} \mathrm{transp}_q(\mathcal{N}_q, q') \right) \right) \\
&\qquad \textcircled{A} \left(\underset{t \in \mu_C(Triple(q_4))}{\textcircled{I}} \left(\underset{q' | \mathcal{N}_q^t . \, \mathrm{required}(q')}{\textcircled{Q}} \mathrm{transp}_q(\mathcal{N}_q, q') \right) \right) \Bigg) \\
&= \mathrm{accessible}(q_4) \circledast \Bigg(\left(\underset{q' | \mathcal{N}_q^{t_1} . \, \mathrm{required}(q')}{\textcircled{Q}} \mathrm{transp}_q(\mathcal{N}_q, q') \right) \\
&\qquad \textcircled{A} \left(\underset{q' | \mathcal{N}_q^{t_2} . \, \mathrm{required}(q')}{\textcircled{Q}} \mathrm{transp}_q(\mathcal{N}_q, q') \right) \\
&\qquad \textcircled{A} \left(\underset{q' | \mathcal{N}_q^{t_3} . \, \mathrm{required}(q')}{\textcircled{Q}} \mathrm{transp}_q(\mathcal{N}_q, q') \right) \Bigg) \\
&= \mathrm{accessible}(q_4) \circledast \Big(\mathrm{transp}_q(\mathcal{N}_q, q_{f-s}(t_1)) \ \textcircled{A} \ \mathrm{transp}_q(\mathcal{N}_q, q_{f-s}(t_2)) \ \textcircled{A} \ \mathrm{transp}_q(\mathcal{N}_q, q_{f-s}(t_3)) \Big) \\
&= \mathrm{accessible}(q_4) \circledast \Big(\mathrm{accessible}(q_{f-s}(t_1)) \ \textcircled{A} \ \mathrm{accessible}(q_{f-s}(t_2)) \ \textcircled{A} \ \mathrm{accessible}(q_{f-s}(t_3)) \Big)
\end{aligned}
$$

For instance, we may choose *average* as the two aggregators, then we obtain the following transparency. We see that the accessibility of the main query q is relatively more important that each of the complementary queries.

$$
\begin{aligned}
\mathrm{transp}_q(\mathcal{N}_q, q_4) = \mathrm{avg} \big(& \mathrm{accessible}(q_4), \\
& \mathrm{avg} \big(\mathrm{accessible}(q_{f-s}(t_1)), \mathrm{accessible}(q_{f-s}(t_2)), \mathrm{accessible}(q_{f-s}(t_3)) \big) \big)
\end{aligned}
$$

Then we deduce the transparency of Wikidata w.r.t. our need.

$$
\begin{aligned}
\mathrm{transp}_q(\mathcal{N}_q, \omega) &= \underset{q | \mathcal{N}_q . \, \mathrm{required}(q)}{\circledast} \big(\mathrm{transp}_q(\mathcal{N}_q, \omega, q) \big) \\
&= \circledast \big(\mathrm{transp}_q(\mathcal{N}_q, \omega, q_1), \mathrm{transp}_q(\mathcal{N}_q, \omega, q_2), \mathrm{transp}_q(\mathcal{N}_q, \omega, q_3), \mathrm{transp}_q(\mathcal{N}_q, \omega, q_4) \big) \\
&= \mathrm{accessible}(\omega, q_1) \circledast \mathrm{accessible}(\omega, q_2) \circledast \mathrm{accessible}(\omega, q_3) \\
&\quad \circledast \Big(\mathrm{accessible}(\omega, q_4) \\
&\qquad \circledast \Big(\mathrm{accessible}(\omega, q_{f-s}(t_1)) \, \textcircled{A} \, \mathrm{accessible}(\omega, q_{f-s}(t_2)) \, \textcircled{A} \, \mathrm{accessible}(\omega, q_{f-s}(t_3)) \Big) \Big)
\end{aligned}
$$

With *average* as the two aggregators, we obtain the following transparency:

$$\begin{aligned}\text{transp}_q(\mathcal{N}_q, \omega) = \text{avg}\Bigg(&\text{accessible}(\omega, q_1), \text{accessible}(\omega, q_2), \text{accessible}(\omega, q_3),\\ &\text{avg}\Bigg(\text{accessible}(\omega, q_4), \text{avg}\Big(\text{accessible}(\omega, q_{f-s}(t_1)),\\ &\qquad\text{accessible}(\omega, q_{f-s}(t_2)),\\ &\qquad\text{accessible}(\omega, q_{f-s}(t_3))\Big)\Bigg)\Bigg)\end{aligned}$$

5 Measures of transparency: first considerations

In this section, we present an introduction to some measures to use in order to quantify transparency (always with regard to an observation need) of a knowledge graph. As seen before, in order to measure transparency of a source, it is necessary to measure accessibility of the query in the source.

5.1 Measuring accessibility

We will describe the measure of accessibility as an example of what can be done. The accessibility is decomposed into two parts, one focusing on the accessibility of the source itself, and the other focusing on the accessibility of the query in the source.

First, measuring the accessibility of a knowledge graph has already been studied [HP20; Fär+18; Zav+16]. The previous definition of accessibility relies only on the protocol of access/ . For a knowledge graph, there are mainly two possibilities. A public SPARQL endpoint is required most of the time, as well as the possibility to download a dump, an RDF export of the graph. The former is seen as more important and the latter is then used by default, or when the constraints on the SPARQL endpoint are too important (execution timeout for the query, restrictions on the queries, ...). We arrive at the following measure, the values are given on an indicative basis.

$$m_{accessibility,endpoint}(\omega) = \begin{cases}1 & \text{if a SPARQL endpoint is publicly available}\\ 0 & \text{otherwise}\end{cases}$$

$$m_{accessibility,dump}(\omega) = \begin{cases}1 & \text{if a RDF dump is publicly available}\\ 0 & \text{otherwise}\end{cases}$$

Hence, the measure of accessibility of the knowledge graph is an aggregation of this two measures. A simple example can be to take the average, but it is also possible to weighted it, according to the importance given to each part.

$$m_{accessibility,source}(\omega) = \frac{1}{2}\left(m_{accessibility,endpoint}(\omega) + m_{accessibility,dump}(\omega)\right)$$

Then, the accessibility of the query in the source relies on three points: the accessibility of the source, a non-empty result, and a reasonable execution time. The first point was presented just before. The second one is easily measured, it consists in executing the query and checking

whether the result obtained is empty or not (see Property 4.1). If the query could not be executed, then the result is empty.

$$m_{accessibility,non-empty}(\omega, q) = \begin{cases} 1 & \text{if the result of } q \text{ executed on } \omega \text{ is not empty} \\ 0 & \text{otherwise} \end{cases}$$

The last point can be measured as follows, with a given θ and for a sufficient number of trials:

$$m_{accessibility,time}(\omega, q) = \frac{\text{Nb of successful trials to get a complete answer for } q \text{ in } \omega \text{ within } \theta}{\text{Nb of trials}}$$

Finally, the measure of accessibility is:

$$m_{accessibility}(\omega, q) = \frac{1}{3}\left(m_{accessibility,source}(\omega) + m_{accessibility,non-empty}(\omega, q) + m_{accessibility,time}(\omega, q)\right)$$

As the accessibility of a source $m_{accessibility,source}(\omega)$ does not depend on the query, it will not be measured each time the accessibility of a query $m_{accessibility}(\omega, q)$ is measured. It can be done only once for the whole set of measures associated with the transparency of a source regarding a need.

5.2 Specifying needs

These measures of accessibility allow to measure transparency of a source with regards to a need already defined. But the difficulty relies on the definition of such need. There is a risk of either being too lax and declaring a source transparent without extensive checks or, on the contrary, declaring any source non-transparent going too far requiring it to meet any conceivable need. The right balance can be difficult to achieve. To evaluate the transparency of a source in absolute terms, the first problem is therefore to determine what needs it should meet. In this section, we discuss several approaches to specify needs.

Adopting the vision presented in Figure 1, users define a need regarding a specific context, either accountability, or reproducibility, privacy, verifiability, or other ones. Each of the contexts of interest is analyzed separately to define questions or queries. This can be done by relying on existing studies on the requirements associated with each context. For instance, a workload has been created to assess the accountability of dataset [OH20] based on three different real-world sources: experts, the Federal Trade Commission, with a view on transparency, and the GDPR (General Data Protection Regulation).

Then, there are possibly two ways to evaluate transparency of the source. Either it can be the aggregation of its transparency evaluated separately on each contextual needs of interest. Then it would be easy to know on which context the knowledge graph is poorly transparent and has to do better, and on which one it is good.

The problem of this method is that two contexts may have several queries in common, especially reproducibility and verifiability for instance. Then, when evaluating transparency by aggregating on each context, some measures are done several times. Hence, a generic need could be constituted by the queries appearing several times in the different contexts: these are the most commonly required queries. Transparency of a source could therefore be evaluated first or only against this generic need. However, this need would be less accurate than the other contextual ones, and the result would be more difficult to analyze.

Another way to specify a need is to focus on the most frequent information required or on the most frequent queries executed. In order to determine them, one possibility is to analyze

SPARQL query logs.The advantage of this method is to analyse what people are generally looking for and to be closer to their need. A difficulty will be to differentiate which queries come under transparency and which do not. Another drawback is self-censorship: some queries might not be executed by people, even if they would like an answer, because there is no chance of success, that they know the graph does not have this information. After extracting frequent queries, or frequent triple patterns, one could put them all in one need, or classify them into several contextual needs.

In the search for frequent needs or contextual needs, some requirements may appear more often than others. For instance, provenance is a very common requirement. For those, it is possible to define new needs, or also to directly define new measures as indicator to look at with the measure of transparency. For instance, one could define a measure of provenance as the proportion of triples in the knowledge graph for which there is provenance information.

Or one could also be looking for specific information every time it encounters a specific kind of IRIs. For instance, in all triples obtained as part of the solution of a query, one could ask that if the subject is a person, then accessing several contact information is needed, as her/his first name and last name and either a phone number, or a physical address, or an e-mail address. These kinds of conditions could be added to build semi-automatically a need: according to the definition 4.9, queries looking for contact information will be required for every need focused on triples having a person as its subject. This also illustrates the recursive aspect of the need.

If a need is defined to evaluate transparency of several knowledge graphs (separately) then it is frequent that the knowledge graphs do not share the same vocabularies. If this is the case, then the need must be defined thanks to precise questions in natural language. Then, either they are translated into queries specifically for each knowledge graph. Or they are first translated into queries using common or recommended vocabularies (PROV-O, VoID, schema.org...) and then they are translated into queries matching the knowledge graphs using owl:sameAs property. This translation could be done semi-automatically, depending on the knowledge graph. A final possibility, probably not the best one, is to consider that the graph should use some specific vocabularies and then to use the queries using common vocabularies as is for all knowledge graphs, without translating them.

6 Conclusion

While transparency does exist in the literature, it appears to be very contextual. Its definition depends on the subject on which it is applied and it seems that absolute transparency, or transparency in general is an unattainable ideal. Hence, we suggest a new definition of transparency: Transparency aims at giving access to the suitable information regarding the needs of observation to any observer.

There also exists numerous concepts around transparency. For instance, reproducibility, accountability, privacy rights, and more, need transparency and all bring specific requirements about which information is suitable: they are particular contexts. Other concepts are necessary for transparency, as for accessibility and openness. Indeed, to reach it, the open and accessible information should be complete regarding the need of the observer.

These considerations allow us to define a new formalization of transparency based on a need of observation. This need holds the contextual aspect. So, transparency with regard to a need is the aggregation of the accessibility of all information items required by the need. The aggregation can be refined to give more importance to certain information, the essential information, according to

the need. It is also possible to structure the aggregation according to a need based on questions and answers instead of information.

When adapting our model to knowledge graphs, questions are naturally transformed into queries. Thus, oracle functions are no longer necessary, there is only a user who defines which are the queries of interest: the user defines the need. This is still a difficult task, but we have already mentioned several directions to explore. Either define some contextual needs or needs based on frequent queries... The structure of the need may be built semi-automatically with some queries to execute each time a certain type of entity appears in the result of the executed queries.

In a following report, we will focus on specifying some generic and contextual needs, enough general to be used as a reference to evaluate transparency of knowledge graphs.

Contents

Bibliography

[Wei+08] Daniel J Weitzner et al. "Information accountability". In: *Communications of the ACM* 51.6 (2008), pp. 82–87.

[PAG09] Jorge Pérez, Marcelo Arenas, and Claudio Gutierrez. "Semantics and complexity of SPARQL". In: *ACM Transactions on Database Systems (TODS)* 34.3 (2009), pp. 1–45.

[GNT10] Jeremy Goecks, Anton Nekrutenko, and James Taylor. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences". In: *Genome biology* 11.8 (2010), pp. 1–13.

[HV11] Noor Huijboom and Tijs Van den Broek. "Open data: an international comparison of strategies". In: *European journal of ePractice* 12.1 (2011), pp. 4–16.

[GM13] Paul Groth and Luc Moreau. "PROV-overview. An overview of the PROV family of documents". In: (2013).

[HSP13] Steve Harris, Andy Seaborne, and Eric Prud'hommeaux. "SPARQL 1.1 query language". In: *W3C recommendation* 21.10 (2013), p. 778.

[MB13] Greg Michener and Katherine Bersch. "Identifying transparency". In: *Information Polity* 18.3 (2013), pp. 233–242.

[Geo14] Nature Geoscience. "Towards transparency". In: *Nature Geoscience* 7.11 (2014), p. 777.

[MJ15] Ricardo Matheus and Marijn Janssen. "Transparency dimensions of big and open linked data". In: *Conference on e-Business, e-Services and e-Society*. Springer. 2015, pp. 236–246.

[Mor+15] Luc Moreau et al. "The rationale of PROV". In: *Journal of Web Semantics* 35 (2015), pp. 235–257.

[GSB16] Alban Gaignard, Hala Skaf-Molli, and Audrey Bihouée. "From Scientific Workflow Patterns to 5-star Linked Open Data". In: *8th USENIX Workshop on the Theory and Practice of Provenance, TaPP 2016, Washington, D.C., USA, June 8-9, 2016*. Ed. by Sarah Cohen Boulakia. USENIX Association, 2016. URL: `https://www.usenix.org/conference/tapp16/workshop-program/presentation/gaignard`.

[Mer16] Meriam-Webster. *Transparent — Definition of Transparent by Merriam-Webster*. 2016. URL: `http://www.merriam-webster.com/dictionary/transparent` (visited on 04/08/2021).

[Wil+16] Mark D Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3.1 (2016), pp. 1–9.

[Zav+16] Amrapali Zaveri et al. "Quality assessment for linked data: A survey". In: *Semantic Web* 7.1 (2016), pp. 63–93.

[Bou+17] Sarah Cohen Boulakia et al. "Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities". In: *Future Gener. Comput. Syst.* 75 (2017), pp. 284–298. DOI: `10.1016/j.future.2017.01.012`. URL: `https://doi.org/10.1016/j.future.2017.01.012`.

[HDL17] Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. "A survey on provenance: What for? What form? What from?" In: *The VLDB Journal* 26.6 (2017), pp. 881–906.

[ABV18] Serge Abiteboul, Pierre Bourhis, and Victor Vianu. "Explanations and transparency in collaborative workflows". In: *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. 2018, pp. 409–424.

[BFC18] Petter Bae Brandtzaeg, Asbjørn Følstad, and Maria Ángeles Chaparro Domínguez. "How journalists and social media users perceive online fact-checking and verification services". In: *Journalism practice* 12.9 (2018), pp. 1109–1129.

[Fär+18] Michael Färber et al. "Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago". In: *Semantic Web* 9.1 (2018), pp. 77–129.

[Mal+18] Stanislav Malyshev et al. "Getting the most out of wikidata: Semantic technology usage in wikipedia's knowledge graph". In: *International Semantic Web Conference*. Springer. 2018, pp. 376–394.

[SR18] Franklin Sayre and Amy Riegelman. "The reproducibility crisis and academic libraries". In: *College & Research Libraries* 79.1 (2018), p. 2.

[Wya18] Daniel Wyatt. "The Many Dimensions of Transparency: A Literature Review". In: *Helsinki Legal Studies Research Paper* 53 (2018).

[Ber+19] Elisa Bertino et al. "Redefining data transparency: A multidimensional approach". In: *Computer* 52.1 (2019), pp. 16–26.

[FTT19] Donatella Firmani, Letizia Tanca, and Riccardo Torlone. "Ethical dimensions for data quality". In: *Journal of Data and Information Quality (JDIQ)* 12.1 (2019), pp. 1–5.

[Sen19] Pierre Senellart. "Provenance in Databases: Principles and Applications". In: *Reasoning Web Summer School (RW'19)*. 2019, pp. 104–109.

[ACM20] ACM. *Artifact Review and Badging - Current*. Aug. 2020. URL: `https://www.acm.org/publications/policies/artifact-review-and-badging-current` (visited on 03/19/2021).

[Ber20] Elisa Bertino. "The Quest for Data Transparency". In: *IEEE Security & Privacy* 18.3 (2020), pp. 67–68.

[Hai+20] Benjamin Haibe-Kains et al. "Transparency and reproducibility in artificial intelligence". In: *Nature* 586.7829 (2020), E14–E16.

[HP20] Elwin Huaman and Oleksandra Paniasuk. *Definition of an Assessment Framework - MindLab D21*. Tech. rep. UIKB, 2020. URL: `https://drive.google.com/file/d/1YCwtuhsAbJUVw1WCaKJejDXXosyQgn4s/view`.

[OH20] Sarah Oppold and Melanie Herschel. "Accountable Data Analytics Start with Accountable Data: The LiQuID Metadata Model." In: *ER Forum/Posters/Demos*. 2020, pp. 59–72.

[SLC20] Daniel Schwabe, Carlos Laufer, and Pompeu Casanovas. "Knowledge Graphs: Trust, Privacy, and Transparency from a Legal Governance Approach". In: *Law in Context. A Socio-legal Journal* 37.1 (2020), pp. 1–19.

[Wik] Wikipedia. *Intelligence assessment - Wikipedia*. URL: `https://en.wikipedia.org/wiki/Intelligence%5C_assessment` (visited on 12/13/2021).