QUALIFIED DATA EXCHANGE

An Introduction

Januari 2023

TNO innovation for life

Rieks Joosten

Inhoudsopgave

1 Introduction

In order to gain access to more/better information, which may come from different sources, parties will engage in data sharing.[1] Such information is meant to provide them with more (or better) insights, enable them to make better decisions, reduce risks, act more adequately, work more efficiently/effectively, innovate, and so on.

The other side of data sharing is where parties make such information available. They may do this because there is a compelling social interest involved, because laws and regulations require it, or because there's a business case.[2]

Here is an example from the Netherlands, where the Municipal Debt Assistance Act (Wgs) obliges all municipalities in the Netherlands to be able to receive signals (data) about payment arrears from 'critical service providers', such as water companies, energy suppliers, housing associates, health insurers, etc. In order to fulfill their legal obligation to care for their residents, municipalities use such signals to decide whether to invite them for a conversation about debts, and the support the municipality or others can offer. The idea is that this can help to prevent them sliding into the situation where the financial problems are too big.

To organize this, all 'fixed costs partners' (water companies, energy suppliers, housing associations, etc.) must be able to share data with the municipalities in which their clients/customers live. The 'network' in which such data is shared is very large and complex.

Figure 1 (on the next page) gives an idea of the complexity of such a network. It shows that each of the (more than 300) individual municipalities can receive signals from one or more water supply companies, housing associations or other landlords, energy and/or heat suppliers, etc. In the Netherlands, there are more than 300 parties that receive such signals (must) provide.[3] That adds up to hundreds of relationships. What the figure does not show, however, is that there is also a dynamic complexity: over time, signal providers will be added, municipalities can be merged, and so on. What the figure also does not show is that there are many more situations in which data must be shared between these and other parties, or even better: between different departments or business units. Each of them has its own data needs and establishes (and maintains) relationships with various other parties. And because these departments also must comply with other legal and other regulatory frameworks for different tasks, the complexity on the work floor is even greater than one would expect at first sight.

[1] Digicampus (2020). The Government as a partner in data sharing . See also: NLAIC, data sharing building block .

[2] NVB, bank-and-data .

[3] there are at least 30+ energy suppliers, 10 water companies, over 300 housing associations (there are probably more landlords), 11 health insurers (each with one or more labels). Source: Internet

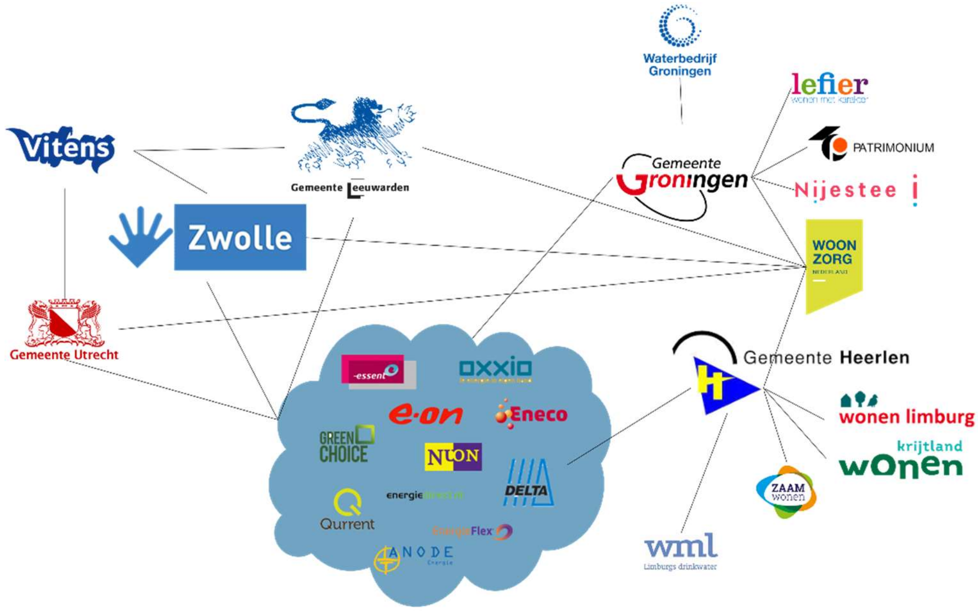

Figure 1: Some relationships between parties that exist to comply with the Wgs.

1.1 Get a grip on complexity

Wanting to have a picture or overview of (departments of) organizations, their mutual relationships, the data they exchange, etc., is often motivated by the desire to get a grip on this. After all, then it can be explained why data is requested, what it is, and what is done with it. This, in turn, is necessary to demonstrate that work is being done within the statutory or otherwise mandatory frameworks. Such a picture is usually constructed with a top-down approach: targets are derived from legislation and regulations, which must be assigned to organizations. Tasks in turn consist of sub-tasks that are assigned to organizational units (departments) – sometimes also from other organizations. This continues until they are finally assigned to (functional) 'roles' within an organization, an organizational unit, or individual officers. This detailing, which is a combination of mandating and delegating, involves more than just assigning tasks (packages) to implementers. With these types of assignments, rules are also given within which tasks must be performed, and what is and what is not allowed (or must). As an overview shows more of these kinds of details, it becomes more and more complex, until it is no longer an overview, but just a confusing mess. It is never complete, consistent and/or coherent, up-to-date, and therefore not really useful. The desire to get a grip on that situation does not become reality.

We also see that in some domains (such as healthcare or information security) frameworks are being drawn up, or compliance with standards (such as ISO 9001 or ISO 27001) is made mandatory. This is also often done to get a better grip on the complexity of this type of data exchange. These high-over frameworks are often further elaborated in handbooks or other tools intended for use in the workplace. And although it often contains very useful things, we see that in the workplace the amount of such frameworks and tools is also quite large. In order to keep things manageable, it must be decided on the work floor which, or which of them, to use (and which not).

There are no clear criteria to determine this, so one chooses what is workable and what is defensible. Whether that also means that people have a grip on the complexity is very much the question.

1.2 Another starting point

We propose to choose a different starting point, one that is based on realizing concrete results on the shop floor, of which it is known who will use them and what they must be suitable for. In black and white terms, this would mean that activities are only carried out on the shop floor if the implementers know which concrete results must be delivered that are 'fit for purpose', i.e., suitable for use by the users of those results. In practice, the soup is not eaten as hot as it is served, but the reverse – simply doing what one thinks is necessary without thinking too much about it – can easily lead to routines that are ultimately inefficient or otherwise undesirable.

By way of example, let us imagine a municipal official whose duties include making decisions about whether to invite citizens for a discussion about debt assistance, as specified in the Wgs. To perform this task (and to be able to make such decisions) he needs data. Each municipality has autonomy in determining the data basis upon which their officials make such decisions. The choices they make have consequences: wrong choices can lead to complaints (for example about violating someone's privacy, or being treated unequally), or making the wrong decision (not inviting a resident that would qualify and benefit).

The autonomy that parties have when it comes to making decisions and collecting data for them means (in theory) that they can make this very simple (for example by throwing dice), or very complicated (so that they can handle everything that could possibly go wrong). In practice, parties typically strive for the simplest way of gathering data that covers the most important risks they perceive to run. Information that is requested therefore serves on the one hand to determine what exactly is being decided, and on the other hand to cover risks. Because risks are always changing (for example, complying with ever-changing laws and regulations), the data requested for a certain decision will also change from time to time.

An important risk that needs to be covered is that of invalid data. A piece of data is 'valid' not only if it has the right meaning (in IT this is about 'syntax and semantics'), but also if its truth can be relied upon, such that when it is used to reach some decision, that decision is valid. This confidence does not have to be absolute: if the risks associated with making an invalid decision are sufficiently low, it is ok to use the data even if it isn’t true. We see this, for example, when buying alcoholic beverages online: the user must tick that he is a holder over the age of 18. The meaning of such a tick is clear, but whether its truth can really be relied upon is highly questionable. But for as long as it does not have any adverse consequences for the web shop, it does not matter. That is very different with a mortgage loan. If the mortgage lender does not properly verify the truth of the applicant being over 18, the transaction may be declared legally invalid with all the adverse consequences that this entails.

To come to grips with the complexity doesn't mean that we need an overview over all data exchanges that might be done, or that every possible exchange of data would need to be facilitated. It is sufficient that every party that runs activities that need data, knows what activities he runs, and for each of them knows what the result of that activity is, who will use that result, and what the result will be used for. That allows him to assess the risks, and determine which data qualifies as ‘valid’ and can therefore be used in the process. This applies to collecting data just as much as to providing data, making decisions, or participating in meetings, and so on. This can be monitored, evaluated and, if necessary, corrected 'decentrally'.

To determine what kind of data is required for a certain task, the starting point is what the operational actors, on the shop floor, need in order to perform that task. From this starting point it can then be determined whether such data are available, where they could come from and ‘certainties’ or ‘qualifications’ they should come with that make them 'valid' to be used in that task.

1.3 Organizing work from this starting point

This change of perspective (from top-down overviews to 'decentralized' bottom-up looking at what is needed) allows organizations to achieve their goals by explicitly linking them to concrete activities with results that are suitable for doing what they intended are.

Suppose a municipality aims to "comply with the Wgs regulation". This means that one or more civil servants are given the task of collecting data that it can be used to determine which residents of that municipality may be eligible for debt assistance, and to invite them for an initial discussion about this. The data must be 'valid' for that purpose.[4]

The municipality must start by determining what these data are, where they must or may come from, and how to assess whether they are valid for deciding whether or not to invite a resident. Different municipalities make different determinations, not only because housing associations, water companies, etc. differ between municipalities, but also because there are (unique) local initiatives[5] that could provide useful data, but that do not exist in every municipality.

Next, the municipality must make such determinations actionable. This consists of choosing the communication channels that may (or must) be used for data collection and determining the set of 'actors' (people or devices) that request, collect and validate the data through these channels. Paying attention to communication channels[6] is necessary, because each type of channel has its own mechanisms for providing assurances that can help determine the validity of the data. Sending data by post provides little certainty about, for example, the sender or its actuality. If the same data were sent electronically, more certainties can be associated with it: for example, a sender can be determined with more certainty if the data is accompanied by a digital signature, or if the data is sent via an SSL connection. The certainties that belong to a certain communication channel (and associated working method) can be provided by technical measures, but can also be of a legal nature, or follow from a system of agreements to which the communicating parties have committed themselves.

Once the communication channels and actors through which data can be obtained have been determined, it may also be necessary to establish 'policies' for obtaining and validating the data. By a 'policy' we mean a collection of rules, work instructions and/or other guidelines intended for a specific type of actors (employees of a certain department or with a certain position, or certain IT systems) and for the performing a specific task (i.e.: a coherent set of actions). The idea is that when such an actor performs (part of) such a task, the actor has this policy at his disposal, he can read it

[4] And, to keep it clear, you also must set priorities. Not every signal may be equally important, and some data may be easier to obtain than others.

[5] Example: 'WE-teams' can only be found (on the Internet) in Groningen and Eindhoven .

[6] A communication channel is understood to mean the entirety of resources and activities that are (or can be) used to send data from a sender to an addressee, and to exchange metadata about it. The latter concerns, for example, requesting/sending a confirmation of receipt, or proof that the data is still valid (not revoked or revoked).

and (correctly) interpret it, and thus can perform the task in the manner specified by the organization is intended.

Finally, the communication channels themselves will have to be made (and kept) available, and it will have to be ensured that there are sufficient human and non-human actors to do the actual work. Naturally, individual communication channels and actors can be used for more than one task. Creating and maintaining such a map can help to efficiently deploy the people and resources that organizations use to do the work that leads to achieving their goals.

The text above, written from the position of the data processor, also applies to setting up the work for data providers. An organization must also determine which (types of) data it wants to be able to provide and make an offer for this. Such an offer not only describes the syntax and semantics of the data itself, but also other properties (which these are, of course, the organization can determine itself). It can be a description of how the data was created (e.g., via a KYC process, or as a result of a certified (technical, or administrative) process), what qualifications the executor(s) had, and so on. These kinds of descriptions are necessary for other parties when they are going to determine whether they will be able to use this data (and whether it is valid) for their purpose(s).

In addition, the organization will have to determine (and describe in an offer) practical matters, such as through which communication channels the data will be made available (and at which 'address' of such a channel this will happen), and which conditions must be met to process a request for the provision of such data. This enables its own organization, as well as other parties that want to be able to request such data, to make the policies for the actors who will do the associated operational work, and to organize that there are sufficient (and sufficiently qualified) people and resources to perform these tasks.

1.4 Look with a different pair of glasses

We can describe this operational perspective even more strictly (more formally). This can assist architects, process designers, etc. in advising/supporting management when they are faced with the task of determining which data is needed for which of their purposes, how to operationally determine its validity, through which communication channels data can be obtained and/or delivered, which actors will be assigned this task and which qualifications they must meet. A more formal description makes it possible to make concrete lists of requirements, wishes, 'mappings', etc. that can facilitate the actual work within an organization. By comparing such lists from different parties later, patterns can perhaps be found that are useful for setting up (more generic) facilities that are not separate from what is needed in operational contexts.[7]

However, designing, implementing and managing information processes and associated data exchanges from the operational perspective does not only have advantages. It can also cause (sometimes intense) feelings of discomfort, insecurity, despondency, anxiety, etc. among designers and implementers. This is not only about a different approach, but also about a different way of thinking. It is like putting on a new pair of glasses, where the 'old pair of glasses' is the current way of thinking, and the new pair of glasses is the model of thinking that we will summarize in the next chapter[8]. 'Putting on new glasses' means that you first take off the old glasses, i.e., that you have to put aside the way you are used to looking at data sharing

[7] With a top-down approach, we often see a tendency to set up 'one size fits all' solutions, which can cause quite a bit of suffering in the workplace – especially when people are involved here.
[8] The eSSIF-Lab Parties, Actors and Actors model has more details on this.

temporarily(!) and this new mental model must actually go to use. That takes getting used to. But the habituation effects disappear once you get used to the new way of thinking and understand how it works.[9] And then you can weigh up the pros and cons and decide whether or when to use these 'new glasses'.

The mental model about data sharing that we describe below consists of a piece of terminology[10] and a description of their interdependence. Words in bold in this document are terms we explicitly define here.[11]

An **entity** (i.e.: something that exists) is called a **party** if it sets its own goals, maintains its own knowledge, uses that knowledge to achieve those goals, and all in an autonomous (a sovereign) way. These are typical people and organizations (or organizational units). We call an entity an **actor** if it can perform actions. Typical examples are people and machines, such as computers. An **act** (or **action**) is any unit of work performed within a certain context by one actor on behalf of one party as a single (indivisible) operation[12]. An example is signing a letter. We say that while an actor is performing an act on behalf of a party, it does so in the role of **agent** (for that party), and that party fulfills the role of **principal** for that actor.[13]

Organizations (or organizational units) cannot perform actions and therefore do not qualify as actors. For example, TNO cannot sign a contract or hire an employee - for this TNO needs an actor who performs these kinds of actions on behalf of TNO, such as a person who is a member of the Executive Board, or someone who has the role of HR employee within TNO. It is nevertheless acceptable to continue using phrases that state that some organization performs an action (as in: "TNO has hired 5 employees today"). However, we *must* then realize that the *actual* meaning of this is phrase is that there is an actor that performs this action on behalf of the organization[14].

People qualify not only as actors (after all, they can perform actions) but also as parties (after all, they have their own objectives, maintain their own knowledge, and so on). If a person performs an action, he/she can do so on behalf of himself (he is his own agent and principal), but also (in the role of agent) on behalf of another party, for example his employer.[15]

Because every action is performed on behalf of one party, that party also determines the rules according to which an actor must perform that action. These are laid down in (detailed) work instructions, (high-over) policy rules, and other

[9] Habituation effects when changing perspective (theories) are well known in history, for example in the transition from a geocentric to heliocentric world view, from Newtonian to relativistic and quantum mechanics, and still are today, for example in understanding (visual) illusions.
[10] We do this by providing a criteria for each term that the reader can use to determine whether something qualifies as (instance of) that term, in order to minimize misunderstandings.
[11] These, and related terms, are detailed in the eSSIF-Lab framework (and glossary) .
[12] See: " Praktijkboek voor Processarchitecten ", van Gorcum, 2002. This defines actions as the basic blocks from which processes and procedures are composed.
[13] An actor can perform different actions for different parties in a certain period, and thus fulfill the role of 'agent' for different parties. However, the actor who performs a single action does so as an agent of (exactly) one party.
[14] The process in which a party is given the opportunity to deploy a certain (whether or not human) actor to (be able to) perform certain actions is called ' onboarding '.
[15] We reserve the term 'employee' (of a party) for an actor who has been onboarded by that party, i.e., that actor has been granted the right (or duty) by that party to perform certain kinds of actions on behalf of that party and has been enabled to do so by this party.

types of artifacts that we generally call ' **policies** ', the content of which belongs to the knowledge of the party that establishes them. For example, policies deal with how a certain decision should be made, what data is needed for it, when something is 'true', under what conditions that data is **valid,** i.e., leads to a valid/correct decision, and so on.

Actors are expected to know and (be able to) follow the policies that apply to the kinds of actions that they are allowed or obliged to perform. This means that there must be a match between the phrasing of such policies and the capabilities, rights and duties of such actors. By properly 'onboarding'[16] actors, a party ensures that they can act according to these policies. Conversely, policies will need to be written such that properly onboarded actors (which we will call **employees**[17] of the party[18] that onboarded them) can executed them as intended. When the capabilities of actors that may execute a certain kind of action can be relied upon, the policy may become simple as it need not spell out how such capabilities are to be used.

Here is an example: a work instruction on the administration of medication is not about how a pill, drops, a syringe, etc. should be administered when the employees that are tasked with this are trained nurses. Rather, it is about the place, time and circumstances in which this should be done. For untrained employees, the policy for these kinds of actions would include guidance about how to administer pills, drops, a syringe, etc. For technical actors such as (running) hardware- and/or software components, policies would typically come in a machine-readable formats that they can interpret, and may be called configuration files, or settings.

This mental model sees parties as completely autonomous (sovereign) entities when it comes to their knowledge (goals, work regulations, etc.). For example, it is not self-evident that they comply with the law: they choose (consciously or unconsciously) whether, or to what extent, they do so. Anyone who has ever (consciously or unconsciously) chosen to drive too fast, or through a red light, or who remembers reports about organizations that (again) have 'gone wrong' will agree that the model thus models the actual reality.

The autonomy of parties is an important principle that we explicitly consider when it comes to sharing data. After all, it means that each party decides for itself what its mission and other goals are, what and how things are done, which rules are (not) followed, with which other parties they interact, what they do, and how, etcetera.

For data sharing, this means that each party will have to determine for itself – on a technical, organizational and legal level – which data it needs for which specific purpose, from which source(s) such data must come, what their syntax and semantics are, what makes the data reliable, and what other properties they must have to be 'valid' in order to be able and allowed to be used for that specific purpose(s). We will elaborate that later.

[16] When an actor is ' onboarded ' by a party, that party has established the kinds of actions that the actor will be expected to execute, the associated rights and duties. Also, the party will have arranged for providing the actor with the circumstances, resources, etc. that are necessary for it to execute its assigned tasks.

[17] Note that this term will not only be used for human actors that are onboarded. We argue that also technical actors, e.g. web servers, robots etc. can be onboarded, and hence qualify as employees of the party that has done so.

[18] We use the term **employer** (of an actor) to refer to any party that has onboarded the actor.

Finally, we define the term '**role**' (of an entity – typically a party or an actor) as a collection of characteristics that this entity has and/or actions that the entity is allowed to perform and/or pieces of knowledge that the entity has, in a certain context. For example, ‘employer’ is a role that a party performs in relation to an actor that it has onboarded (the actor performs the (associated) role of ‘employee’). Performing this role of ‘employer’ means that the party decides which tasks, rights and duties to assign to its employees. Another role that a party may perform is that of ‘data consumer’. In this role, the party decides what data it needs for what specific purposes, and what criteria the data must fulfill in order to be valid for such purposes. Actors, too, can perform roles. In the role of ‘agent’ (of a party), an actor executes an action on behalf of that party. Performing a particular rol may require that another role is performed as well. For example, an actor cannot be (perform in the role of) an agent for a party unless it is an employee of that party. What (performing in) a role entails should be properly defined by the party that uses the term. Such definitions are crucial for a proper and shared understanding.

It is important not to confuse a 'role' with the party or actor that fulfills such a role (in a certain context). After all, that party may be able to fulfill a different role at a different place or time. Consider a party that offers to provide certain kinds of data. When such a party receives a request to supply such data, it will first perform in the role of data consumer (we will define that role later), gathering the data that it needs to determine whether to accept or reject that request, and then, after the request is accepted, it will perform in the role of data-supplier and provide that data.

When we say that <*role name*> does something, we mean that an entity that at a certain moment fulfills the role <*role name*> does this something. For example, if a data consumer requests data, and later a data provider provides data, these roles may be fulfilled by one and the same party. But it is also possible that these roles are fulfilled by different parties.

In summary: the new lens mainly consists of replacing the terms 'organization' and (the different variants of) 'person' (natural person, legal entity) by the terms 'party' and 'actor', which have a very specific meaning with which we make a different distinction: a party is an autonomous unit that manages its own (subjective) knowledge, and an actor is an entity that can do things. When an actor does something, he always does it on behalf of a party, and that party also provides the knowledge (policies) that the actor uses to perform the action(s) in the way that this party has devised. When we say (yet) that a party performs an action (does something), we always mean that there is a certain actor who performs this action on behalf of this party.

2 Qualified Data Exchange (QDX)

Sharing data involves a lot and this is usually done from the perspective that data that is available somewhere must also be able to be used elsewhere[19]. The current view is that providers of such data properly specify the syntax and semantics of such data ('semantic interoperability') and that all kinds of ('trust') frameworks will help to actually (re)use that data.

To this perspective from the data provisioning side, we will add that of the individual and autonomous data processors (data consumers). From the data consumer perspective, data is obtained and further processed for a specific purpose. Perhaps it serves to make a decision, or to perform action. It is then important that the data must be suitable (valid) for that purpose: after all, a decision based on invalid data can entail undesirable consequences (risks). This point of view emphasizes that a data processor should not only be able to **verify the data it obtains**, i.e., determine that the data has the intended syntax and semantics, and is an authentic and timely statement of its provider, but ALSO be able to **validate**, i.e., determine whether it is valid for the purpose for which it intends to use it[20].

Qualified Data is data that meets all conditions that the party requesting such data has specified in order to establish that the (obtained) data is valid (valid) for (further) processing in order to produce a specific result (achieve a particular purpose). Conditions can relate to syntax and semantics, but also to the way in which the data was created, guarantees regarding origin and/or integrity, etc.

Qualified Data Exchange (**QDX**) is a way of looking at data sharing with the main starting points being the autonomy (sovereignty) of all parties, and the (subjectively derived) matters such as policies for task execution, data needs, validity criteria, etc. QDX is a model that describes the generic viewpoints (roles) for dealing with such things as offering, consuming, and exchanging Qualified Data. Figure 2 (on the next page) shows a (simplified) overview of this.

The right column shows how QDX works from the perspective of a party that needs data to perform certain types of actions (processing). He must specify the kinds of data he needs, and – for each individual purpose – the criteria such data must satisfy in order to be valid for the envisaged kind of processing. Such specifications are then translated into work instructions (which we call ‘policies’) that designated employees can use to determine which communication channel is to be used, how to formulate appropriate requests, how and where to send such requests to, how to receive responses, validate the data therein, and further process them to realize the expected result.

[19] This is mainly viewed from the perspective of data suppliers and/or data ecosystems.

[20] Checking syntax and semantics is necessary, but not sufficient. Depending on the intended purpose, it may also be necessary to know who determined this data (e.g. the government, a bank, a water company), and/or how it was done (e.g. a real measurement, or an estimate), and / whether certain (legal) rights can be derived from the data, etc.

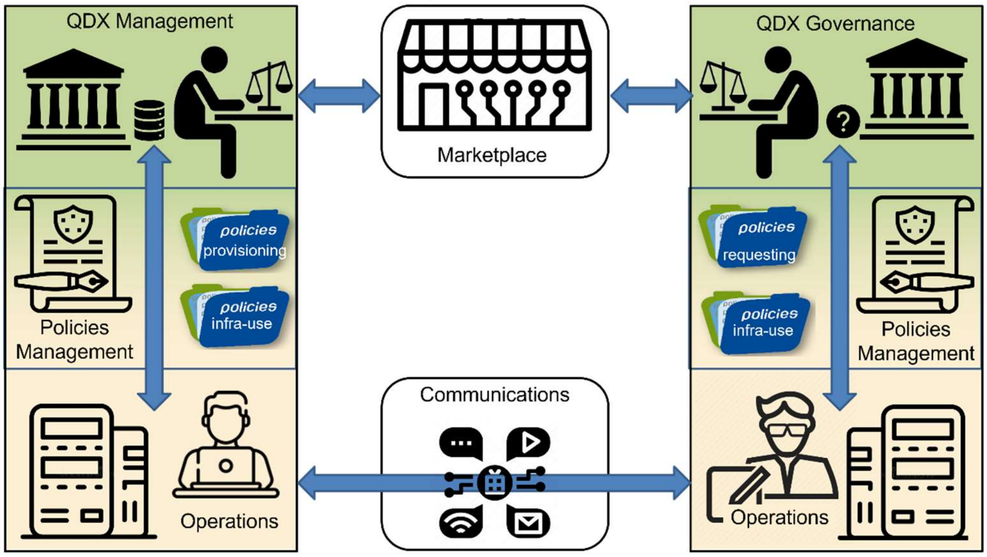

Figure 2 QDX Overview

These policies are created and maintained earlier (design - time[21]) in the 'policies management' process for different operational ('run-time'[22]) contexts[23] (different communication channels[24] and/or types of actors). The QDX governance process determines which processing can be done on behalf of the party, which data is required for this, through which communication channels it may be requested, and when they are valid for a certain processing.

The left column shows how QDX works from the perspective of a party that can provide data. In its QDX management process, it is determined what data that is, what the terms of delivery are, and what metadata will be published that enable parties to determine whether that data would be valid for specific purposes that it might need them for. Again, these decisions must be translatable to the operational environment, i.e., into policies that can be used in the various operational contexts for data delivery by the relevant actors to assess requests for the provision of certain data, and if they are accepted. arrange for the data to be delivered.

At the top in the middle, we see that it must be possible to match demand (from the right column) and supply (from the left column). In fact, this is a functionality that is comparable to a marketplace: the provider must be able to publish its offers, and the user must be able to find the offers he needs from this. It implies that providers must be able to advertise their offer in such a way that processors can determine whether or for which actions they want to be able to perform with it, it is suitable. It

[21] That is: in the preparatory work. This is part of activities associated with (the risk management part of) process design and/or information modelling.
[22] The term 'run-time' refers to the time during which actual operations/actions are performed. This contrasts with 'design-time', which refers to the time in which such operations/actions are specified.
[23] That is: in the operational work, if one (real) decision is taken by one specific actor in one specific situation.
[24] A communication channel is understood to mean the entirety of resources and activities that are (or can be) used to send data from a sender to an addressee, and to exchange metadata about it. The latter concerns, for example, requesting/sending a confirmation of receipt, or proof that the data is still valid (not revoked or revoked).

also means processors need to know where to find those ads so they can set their policies.

At the bottom we see that actors (in the right column) can request data from actors in the left column. They all use the specific policies of the party on whose behalf they perform their respective task. That's what 'data sovereignty' means. At the bottom in the middle, we also see that such actors can use communication services provided by other parties for the actual exchange. This could be, for example, a postal company (for physical exchanges), but also a network supplier (for electronic exchanges), or something else.

In the following chapters we elaborate on some of these blocks.

2.1 QDX Governance

An actor (person or IT system) who performs an action on behalf of a party (run-time[22]that requires data, such as making a decision, must be aware of a policy that shows which data is involved and which validity criteria[25] are involved. belong. This policy must therefore have been specified[26] and established earlier (design-time[27]) and must also have been made available in an adequate manner for every actor who performs actions of that kind. Design-time choices are therefore made and translated into (possibly several[28]) policies that make the run-time exchange of (valid) data possible (or impossible).

By **QDX governance** we mean the design-time process that a party carries out, in which it specifies and determines what kind of data is required for what kind of actions (if these are performed on its behalf), and what validity criteria are involved. In doing so, the party will consider the risks involved in performing that action if data were used that are incorrect, do not relate to the correct 'subject', are too dated, are unreliable (according to the judgment of the data consumer of course), may not be legally used, etc. The result of this is that for each type of action the party specifies the frameworks for the data and collateral that are directly relevant not only for the performance of the action, but also for the validation of data and risk mitigation. These frameworks may also impose restrictions on the ways (communication channels) in which the data is exchanged.

The result of the QDX governance process is a specification of the frameworks for obtaining data that are important to the party for the various processing operations that the party carries out with it in order to achieve its own goals. These frameworks contain all the information needed by the actors performing the 'Policies Management' process to translate them into the rules, work instructions and other artifacts that are then used by operational actors when collecting that data, so that it is done in ways that suit within the frameworks set by the party.

[25] That is: criteria (linked to a specific goal) that can be used by a specific type of actor to determine whether a set of (supplied) data is valid to be used for the relevant goal.

[26] Specifying is writing down the content of the policy (clear, unambiguous, consistent, complete, etc.).

[27] Establishing (of a policy) is deciding that the policy should actually be used.

[28] Different kinds of actors need different kinds of policies. For a human being, a policy must be readable in the language that people understand – for international companies, policies in several languages may be required. IT systems require 'machine readable policies' – files that can be used by these types of systems in such a way that they perform the actions as intended (design-time).

2.2 QDXManagement

A party that has data and wants to share it must ensure that this data is made available in an adequate manner for parties that would/must/be able to use it. The data consumer will also have to disclose details about such an offer ('marketing').

By **QDX management** we mean the process that a party carries out, and in which it determines what (type of) data it wants and will publish (syntax, semantics), under what conditions the data is provided (for example to whom, what they can do with it , etc.), which types of security are offered to customers (for example, regarding the way in which the data came about, the qualifications of those who did so, or regarding its topicality, etc.). Frameworks are also established regarding the operational delivery. For example, restrictions can be imposed that state which communication channels can (not) be used, which customers deliveries may (not) be made, and so on.

The data provider will consider the risks that the provision of this data (in the various modalities) entails, for example to parties who are not allowed to have/use this data, or if the law imposes restrictions on the sharing of that data. He will also check whether there is a business case for him: a fee may be charged for some data; for others it may be a legal requirement.

The result of the QDX management process is a specification of the frameworks that a party sets for the delivery of data. These boxes contain all the information necessary to

- enable the actors performing the 'Policies Management' process to create and manage the rules, work instructions and other artifacts (policies) that are then used by operational actors to determine whether (a) a received request for the provision of data must be declined, and (b) if such a request is granted, where the data may be sought and how it may be provided to the requester. This ensures that data is only supplied in ways that fit within the frameworks set by the party.
- go through the process that results in what we call a **QDX advertisement** (for a certain set of data), ie a document containing all the data that a potential data buyer needs to decide whether to use the advertised data for one or more of his goals, whether they are valid for that purpose, and also whether he can use them to create his own (user) policies. A QDX advertisement therefore not only contains the syntax and semantics of the data offered, but also information about certainties and delivery conditions, the communication channels used, the 'addresses' or 'endpoints' to which requests must be sent, the structure of such requests, etc.

2.3 Policies Management

By 'Policies Management' we mean a process that serves to 'translate' the frameworks established in a governance or management process into operational reality.

The scope of such a process is usually broader than just collecting or supplying data: it is also used for translating management/governance frameworks into other primary and secondary processes. In the figure, however, it is expressly about the frameworks that relate to the supply and/or collection of data.

Looking through the new, more formal lens, we assume that for each processing (supply or collection) of data by a party, it has been determined through which communication channels this takes place, and which (classes of) actors are allowed and able to perform certain tasks in operational work. Such an inventory is not very complicated in itself, but if the number of such processing operations increases, it can be a lot of work to make this explicit, and to keep it up to date.

Other inventories are required, e.g. for:

- **communication channels**, and the properties they have that are relevant for policy making. For example, if data exchanges take place electronically,
    - via a network that falls under the (Dutch) Telecom Act, the confidentiality of the communication is guaranteed;
    - via an SSL connection, then confidentiality is guaranteed and there is also certainty about the party with whom communication takes place;
    - via an IDS connector, then there are also guarantees that the party with whom communication takes place adheres to a certain set of agreements;

  Similarly, non-electronic communication channels (such as sending data by post, or by courier) may have properties that are relevant to know when creating data exchange policies.
- **employees**, and their capabilities[29] that are relevant for policy making. For example, for human actors, job requirements (aggregated in a functional role) may be important, so that a policy can be made that states that actors may only collect personal data if they can do so in accordance with the GDPR[30]. Which (other) properties are important will differ per party.
  Similarly, non-human actors (computers/applications) may have properties that may be important for performing data exchange operations. This will mainly concern the question of whether or to what extent they can handle certain communication channels (technically).

Setting up such inventories properly can be quite a job. After all, it must be determined which properties should be included in the inventories, and that in turn depends on what may be required when writing the policies.

The maintenance of the aforementioned inventories concerns

- the entry or exit of actors or the (temporary) suspension of activities. That is not a lot of work in itself, but it does require a certain discipline from administrators that is not always obvious.
- managing the properties of communication channels resp. actors who must have a place in the inventories. If governance/management processes lead to decisions for which the policy writers need different/new properties, this will have to be documented in the registrations. However, the idea is that this will not happen very often.

[29] There are many ways to organize such an administration. One would be to assign ‘functions’ to employees, where a ‘function’ is assigned to an employee if it is expected to execute certain tasks and has certain capabilities that allow him to do so as its employer intends them to be executed by following the appropriate policy. But there are other ways as well.

[30] General Data Protection Regulation .

With the help of such (well-maintained) inventories, a QDX policy for a certain data delivery or data request can be relatively easily translated to the communication channels and actors who do the operational work. This does not alter the fact that every policy must be suitable for the (type of) actor that has to work with it. If that is an IT application, this policy will probably consist of program code, or as a configuration file. For human actors, this can be a work instruction that is set in a natural language that the person in question has sufficient command of.

The result of the Policies Management process (at least as far as QDX is concerned) is a set of policies, one for each combination of a (type of) actor and a (type of) communication channel, such that these actors perform the tasks associated with the delivery or requesting data for a certain processing operation operationally within the framework of the party on whose behalf they do so.

2.4 QDX marketplace

A QDX marketplace is a (physical or logical) place where parties (in their role as data provider) can advertise their data offerings, and (in their role as data consumers) can check what data is being offered by other parties. What is typical of a QDX marketplace is that a data offering is not only about the type of data (syntax and semantics), but also mentions all sorts of other data that data consumers need to determine whether they can use the data in one or more of their data processing operations – i.e.: whether they are valid for such data processing.

We think that a QDX marketplace is ideally situated within a 'community' (or ecosystem), i.e. within a group of parties that already have something to do with each other. After all, such communities are quite capable of making agreements in a relatively simple manner that offer guarantees that are important for data consumers (but also for data producers) to label data as 'valid' for certain processing operations. Then the simple fact that parties are members of such a community already provides important certainties, which may make obtaining even more certainties superfluous, and thus greatly facilitates the exchange of data.

A QDX marketplace can be effectively realized as a platform and/or (online) catalog containing (all data of) QDX advertisements. Within a community, this place will be easily communicated so that parties can upload their advertisements there and view those of others. These catalogs will (for the time being) at least have to be readable by people, because they must be able to decide with the data whether they want to use this type of data (and therefore also request run-time). We see them as an extension to existing data catalogues, where the extension consists of also specifying the collateral and delivery conditions.

2.5 QDX matching

Under **QDX matching** is understood to mean all actions that parties in their roles as data consumer and data provider perform during design-time in order to match the supply and demand of data. This is akin to 'semantic interoperability', where syntax and semantics of data are aligned, which then leads to a standardized offering that data consumers then (run-time) conform to. But here it is emphatically also about coordinating what the associated certainties are that can or must be supplied, and how a data consumer can verify this in order to validate the 'ordinary' data.

Matching the supply and demand of data is conceptually the same as matching the supply and demand of any other product (or service). There are therefore different ways to shape this. For example, it can be 'remote', where a data provider advertises its data ('products') and waits to see who will buy it, and the data - consumer somehow sees those advertisements (as 'spam', or because he looked for) and then sees what he can use. The aforementioned catalogs can play a useful role in this.

In the context of data sharing, it is a bit more nuanced, because it is not really about the tangible ('tangible') data, but about the intangible ('intangible') information that is represented by this data. Each party is autonomous and determines (independently) which data it uses to represent certain information that it knows. Terms like "reliable", or "true" will mean different things to different parties. And the mapping (i.e.: semantics) that the data provider chooses to use must not only be discoverable by the data consumer, but in addition, the mapped information (the concepts behind it) must fit into that data consumer's mental models. This is called: 'semantic interoperability', and it requires a more intrusive way of coordination than is required, for example, between a supplier and consumer of ordinary products (a radio or TV).

2.6 Operational data exchange

All the foregoing serves to make it possible to exchange data operationally (run-time[22]). More concretely: a party that wants to carry out an action for which certain data is required, in principle has everything it takes to ask for this data, and to determine from the response to those questions whether this is the data that it needs. and also, whether or not they are valid. This applies to all combinations of contexts and actors (people, machines) made possible by such a party (design-time).

The same applies to a party that wants to perform (or have performed) actions that concern the handling of a request for the provision of certain data (the result of which must be whether or not that request is granted), and – if such a request is granted – collect the requested data and send it to the requested destination via the chosen communication channel.

In order to actually be able to request the data required and to receive an answer, at least one communication channel is required that makes this possible. The parties will want to limit the number of communication channels, because each of them entails set-up and management costs. On the other hand, a party will want to make its data available in various ways in order to enable as many parties as possible to also purchase it. Furthermore, each exchange modality has its own unique properties, which can contribute to fulfilling the validity criteria that data consumers (for certain data/purposes) have specified. We have discussed this in more detail before (in 'Policies Management').

3 Reflection

QDX is based on several principles that are not self-evident in practice. For example, we regularly see that when setting up an electronic data collection process, people think that the first step is that 'the user must log in': after all, you have to know who you are dealing with. It is often assumed, tacitly or not, that the user is also the party responsible for the correctness, timeliness, etc. of the data. This applies both to human 'users' and in situations where the data from an IT system comes from another party (where that system is often authenticated by means of a PKI certificate). Such ideas often arise from procedural thinking; one imagines *how* the process proceeds and that is then arranged.

QDX asks you to stay focused on your own goals, the associated results, who will use these results and what they should be able to do with them. It's more about the 'what' than the 'how'. And then you can't help but think about what kind of data is needed for that, where it should come from, and what makes it valid to be used. It is therefore no longer self-evident that the party that actually supplies the data is the same as the party that is the source (the author) of the data. It is then conceivable that citizens collect data about themselves that originate from a multitude of parties (their employer, the tax authorities, banks, etc.), in order to share it when necessary, with parties that need it.

It therefore becomes easier to consider that certain data processing does not actually have to be done by the party that needs the results. This applies especially to (automatable) calculations or reasoning. There are techniques, such as Multi-Party Computation (MPC), in which (parts of) such calculations are performed by parties who have certain data, and only the results are shared. In this way, calculations can be made (elsewhere) that a party could never have done independently, for example because the data used is too sensitive and should therefore never be shared.

QDX makes a clear distinction between parties (entities that maintain their own subjective knowledge and make decisions about what to share or what they need from others) and actors (entities that can do something - that mainly concerns operational questions, supplying and validating data). Making this distinction makes it easier to devise new forms of data delivery and data request. It then no longer matters much whether data comes directly from 'the source', is supplied via an intermediary/intermediary, via a 'safe' (whether or not in the cloud), as long as there are sufficient certainties – for example about its source, its integrity, timeliness, reliability, etc. – which make the data valid for the purpose for which the data consumer wants to process it.

Although QDX makes it quite clear for a single processing what is involved, most parties have to deal with many - often even very many - processing. And what is clear for a single processing operation is no longer so for such a multitude of processing operations. You have to be able to organize well to keep an overview, and that is no sinecure.

But you can also start working together. In practice, this often already happens; there are already many communities whose members work together, know who/how someone can be trusted, etc. This makes us think that these *communities* can also play an important role in organizing the QDX related work.

We call such an (existing) partnership a **QDX community** if it (in addition to its existing goals) also aims to facilitate the idea of QDX for the participating organizations. This can be done, for example, by:

- making an inventory of the types of processing carried out by many of the cooperating parties.
- specifying a (minimum/maximum) data sets (syntax and semantics) that are relevant for carrying out such processing.
- drawing up (minimum/reasonable/maximum) validation criteria for these datasets.
- specify qualifications for these datasets, especially those that facilitate evaluation of validation criteria, and how they are communicated in QDX advertisements.
- setting up a QDX marketplace in which all participating parties can place their QDX advertisements.
- selecting one or more communication channels that can (or must) be used within the collaboration for mutual data exchanges.
- drawing up (minimum/reasonable/maximum) qualification standards for actors who perform tasks in supplying or consuming data.
- writing the policies for the combinations of (agreed) communication channels and actors (who meet certain qualification standards), so that they can easily exchange the 'standardized' types of data.
- setting up a 'trust framework' that fits the needs and capabilities of the members in order to operationally determine as easily as possible which data (suppliers) are reliable, with which parties' data may be shared, etc. – whatever is important for the members of the partnership.
- and so on.

The idea of communities is increasingly used and is then known as a 'data ecosystem' or a 'data space'. This usually concerns the establishment of the governance of an infrastructure for exchanging data between members (and possibly non-members), for which there are already many options, such as IDS, iShare, SSI, Gaia-X, FIWARE, etc.

One of the first (successful) data spaces is the Smart Connected Supplier Network (SCSN)[31], which provides for data exchange between parties in the manufacturing industry. This is an example based on the IDS infrastructure. The European Commission encourages the creation of data spaces, and we expect the numbers to grow significantly in the coming years.[32]

The importance of a focus like QDX on achieving goals/ results, and in particular on establishing the validity of the data needed to achieve them, will only increase. For example, the idea behind 'Explainable AI' is actually exactly that: establishing the validity of AI outcomes for use in specific processing operations. The recently made public ChatGPT , and especially the way in which it was immediately used and abused, only underlines this importance.

[31] See https://smart-connected.nl/nl . Their processes are also available online .

[32] See European Commission – Shaping Europe's digiatl future: Staff working document on data spaces .

We expect that both governments and private, commercial parties will benefit from a way of looking at data sharing that is primarily based on the 'what' (instead of the 'how'), and therefore better focuses on these types of interests and addresses associated concerns. This does not solve all problems, but it does offer a new perspective on data sharing.