Data Journeys: Identifying Social and Technical Barriers to Data Movement in Large, Complex Organisations

DOI:
10.1016/j.jbi.2017.12.001

Document Version
Accepted author manuscript

Data Journeys: Identifying Social and Technical Barriers to Data Movement in Large, Complex Organisations

Iliada Eleftheriou[a,*], Suzanne M. Embury[a], Rebecca Moden[b], Peter Dobinson[b], Andrew Brass[a]

[a]*School of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester, M13 9PL, UK*
[b]*Royal Bolton Hospital, Minerva Rd, Farnworth, Bolton, BL4 0JR, UK*

Abstract

Managers in complex organisations often have to make decisions on whether new software developments are worth undertaking or not. Such decisions are hard to make, especially at an enterprise level. Both costs and risks are regularly underestimated, despite the existence of a plethora of software and systems engineering methodologies aimed at predicting and controlling them. Our objective is to help managers and stakeholders of large, complex organisations (like the National Health Service in the UK) make better informed decisions on the costs and risks of planned new software systems that will reuse or extend their existing information infrastructure.

We analysed case studies describing new software developments undertaken by providers of health care services in the UK, looking for common points of risk and high cost. The results highlighted the movement of data within and between organisations as a key factor. Data movement can be hindered by numerous technical barriers, but also by other challenges arising from social aspects of the organisation. These latter aspects are often harder to predict, and are ignored by many of the more common software engineering methodologies.

In this paper, we propose data journey modelling, a new method aiming to predict places of high cost and risk when existing data needs to move to a new development. The method is lightweight and combines technical and social aspects, but relies only on information that is likely to be already known to key stakeholders, or will be cheap to acquire.

To assess the effectiveness of our method, we conducted a retrospective evaluation in an

NHS Foundation Trust hospital. Using the method, we were able to predict most of the points of high cost/risk that the hospital staff had identified, along with several other possible directions that the staff did not identify for themselves, but agreed could be promising.
Keywords: data movement, information sharing, data journey, socio-technical barriers, data flow

1. Introduction

Technological advances and business changes drive organisations to develop new, more advanced information systems (ISs) to share and integrate their information. But realising value from these new ISs requires hard decisions to be taken. Is the new system development or re-design worth undertaking? Is the value to be gained more than the costs of developing and maintaining the new development?

Predicting costs and risks of software development is hard [1]. There is a complex mix of factors to be considered when deciding whether to proceed with a new development or not. Technical difficulties arise when sharing or integrating information, often stemming from the diverse data sources involved. Other, often neglected challenges stem from the social aspects of the organisation: its people, policies, processes, governance, etc. Examples can be found in the health care domain, in which people have been found to be reluctant to change their current processes to use the new system in place, or user requirements are not met because of conflicting organisational policies and governance issues [2, 3, 4]. Specifically, Greenhalgh *et al.* show the importance of human factors in the integration of electronic patient record (EPR) systems [5]. They state that lack of consideration of human factors is detrimental when bridging the model-reality gap in EPR systems.

Costs arising from technical and social barriers are often underestimated, especially in complex developments [6]. An example is the National Programme for IT (NPfIT), an

*Corresponding author
Email addresses: iliada.eleftheriou@manchester.ac.uk (Iliada Eleftheriou), suzanne.m.embury@manchester.ac.uk (Suzanne M. Embury), rebecca.moden@boltonft.nhs.uk (Rebecca Moden), peter.dobinson@boltonft.nhs.uk (Peter Dobinson), andy.brass@manchester.ac.uk (Andrew Brass)
URL: www.datajourney.org (Iliada Eleftheriou)

initiative by the Department of Health in the United Kingdom (UK) aiming to use modern information technologies to improve the delivery of health services and the quality of patient care [7]. Numerous information sharing and integration solutions were introduced under NPfIT, but after 12 years and a total forecast cost of 9.8 billion UKP, the planned central, integrated system has yet to be established [8, 9, 10]. Another example is the e-borders programme of the Home Office in the UK. The programme was initiated in 2003 and aimed to better control UK borders by integrating information from external stakeholders, like plane, train and ferry carriers. However, costs and effort were underestimated and the programme terminated in 2014 with a total cost of 830 million UKP [11].

Existing approaches to managing risk and estimating cost of software developments do not ease the decision making process. They are principally focused on creating detailed predictions based on substantial models of the planned development [12, 13, 14, 15, 16]. They are aimed at supporting project managers throughout the development process itself, rather than giving a low-cost indicator for use in early-stage decision making.

Despite the plethora of modelling techniques and notations found in the literature for use during information systems design [17, 18, 19, 20, 21, 22, 23, 24, 25, 26], we only found a handful of methodologies that give equal prominence to both social and technical factors [23, 24, 27]. Of these, none were sufficiently lightweight to be used in early stage go / no go decision making.

To address this need, we devised a lightweight method to predict places of high cost and risk when existing data is moved to a new system development. We began by analysing a collection of 18 case studies, written by staff working for the National Health Service (NHS) in the UK. The case studies describe factors that contributed to the failure or success of recent IT developments with which the authors of the studies had been involved. The results of the analysis showed that the IT projects failed due to a mixture of human and technical factors, with the human factors being by far the most dominant. This is consistent with results from other sources (e.g., [28, 29]). However, our analysis also revealed that data movement was both a common feature across the failing projects *and* a socio-technical phenomenon. That is, we observed that problems occurred when data was moved between

groups of people (through sharing audio recordings or paper-based records, for example) as well as when it was moved between systems. This led to the hypothesis that data movement (between people, between systems, and between people and systems) could give us a high-level indicator of both technical and social costs and risks within planned IT developments.

From this hypothesis, we propose a method for early-stage cost/risk prediction based around the creation of *data journey models*: high-level models of the journeys that data makes through networks of people and systems in order to provide some value. In an earlier paper, we described the results of our analysis of the 18 case studies, and how they led to the notion of data journey models [30]. In this paper, we present the data journey modelling method in full, and also describe an evaluation we undertook within an NHS Trust Hospital. Specifically, the contributions of this paper are:

- A lightweight method that combines social and technical information to predict barriers to data movement that can impose high costs and risks in planned IT developments.
- The application of our method to a real world case study from the health care domain, describing data movement from a General Practitioner (GP) organisation to the radiology department of a Foundation Trust (FT).
- A retrospective evaluation of the method to assess the precision of the predictions with the help of NHS domain experts.

The results of the evaluation are promising. Our method was able to predict almost all of the costly elements of the former system that the hospital staff had previously identified. Also, the method identified other possible cost-saving actions that the staff didn't identify, but that NHS domain experts agreed could lead to cost savings.

The following sections describe the process we used to develop our cost/risk prediction method (section 2). We present the fundamental model on which the prediction method is based: the data journey model (section 3) and show how it can be used to predict points of cost and risk (section 4). We describe the evaluation exercise we undertook to test the

efficacy of the approach (section 5) and, finally, draw conclusions and make suggestions for future work (section 6).

2. Methods

We first set out the steps we took to design the data journey modelling method. We were fortunate enough to have access to 18 case studies written by NHS staff taking the "Informatics for Healthcare Systems" professional development course at the University of Manchester, during the 2013/2014 academic year. They describe a variety of IT developments in the NHS, covering cancer care, ambulance service management, in-patient management, heart failure care, diabetes care, bed management and more. The authors of the case studies were asked to categorise the new developments as successful or not, and to identify the human, organisational and technical factors leading towards the success or failure of the system. Only 3 out of the 18 studies were categorised by the authors as having been successful. The rest were described as having (completely or partly) failed to deliver the expected benefits.

The information contained in the case studies was the primary input to the design of our method. We took the following steps in analysing and using the information they contained:

1. We examined the case studies and extracted a list of factors that were present when IT failure of some kind occurred. This was a relatively trivial task in our case, as the authors of the case studies had been explicitly asked to identify and record the factors, social and technical, that contributed to the project outcome. We had only to extract them, to integrate them into a common set and to regularise the terms used.

   The most common failure factors found in the case studies were related to people: for example, staff resistance to process change, insufficient stakeholder engagement in decision making and lack of shared vision. Other factors of both a technical and organisational nature were identified. Examples include conflicting data formats, disconnected data silos, inadequate system performance, governance issues, complex organisational

structures and lack of political influence. In all, we identified 32 failure factors, which are described elsewhere [30].

The most important outcome from this step was the realisation that failure often went in hand with the movement of data outside its familiar context and into some new setting, with new users, new systems and new requirements.

2. Next, we examined the case studies once more to identify the types of data movement involved in each one. This was more challenging, as the authors had not been asked to characterise the data movements involved. We had to infer the presence of movement patterns from the explanation in each case. Again, we had to combine and regularise the patterns from different studies to create a consistent set.

   The outcome of this step was a catalogue of data movement anti-patterns. An anti-pattern is defined by Ambler to be "the description of a common approach to solving a common problem, an approach that in time proves to be wrong or highly ineffective" ([31], p.20). Similar definitions are given by Budgen and Koenig [32, 33]. Though the term is more normally associated with software design, it can also be applied more broadly. In our context, a data movement anti-pattern is a commonly occurring movement of data that appears to solve the problem of conveying needed data to a consumer, but which produces higher-than-expected costs in the longer term, that can reduce the overall value obtained from the project as a whole.

   We were able to synthesise a total of 8 data movement anti-patterns from the case studies, including movement between people, between systems and between people and systems. We found that movement across a discontinuity, whether technical or social, leads to higher costs. For example, if a source system stores data in a physical form, but a new target system requests it in electronic form, then a transformation cost will be imposed on the new system. (This pattern was surprisingly frequent in the case studies.) Similarly, if data is moved between organisations, then information governance or ethical issues may arise. The full catalogue of movement anti-patterns can be found elsewhere [30].

3. Starting from the component elements used in the data movement anti-patterns, we devised a set of modelling elements for representing *data journeys*. A data journey describes how data moves from its place of entry into a software system to the point at which it is used to create value for some stakeholder. The aim was to create an abstract representation of the system that contained exactly the information needed to identify the data movement anti-patterns. The model is able to represent planned (or speculative) journeys as well as existing ones.

4. We then developed a method that uses data journey models to predict the points in the organisations concerned where high cost/risk might occur. The method works by overlaying social and technical constraints (as specified by the movement anti-patterns) onto a data journey model.

The remainder of this paper presents the results from steps 3 and 4 above.

3. Data Journey Modelling

The basis of our prediction method is a model that describes an abstraction of an organisation (or collaborating group of organisations) designed to make explicit the kinds of problematic data movement pattern we identified from the NHS case studies, while ignoring irrelevant details. We call this model a "data journey model", since it shows the high-level journeys data takes through the organisation in order to deliver value.

We use the term *data journey* to describe the movement of one or more data entities through the landscape, from their point of entry to their point of use. Data journeys are purposeful, implying that the data is needed at its destination for some value-creating step. For example, suppose a GP requests a patient blood test from a nearby pathology lab to decide on a further care plan. To do so, data needs to travel from the GP organisation (in the form of a request card and blood sample) to the hospital porter's pigeon holes, to the lab staff, to the lab's database (where results are input by the lab analyst), and back to the GP's database to await discussion with the patient. All these steps together make up

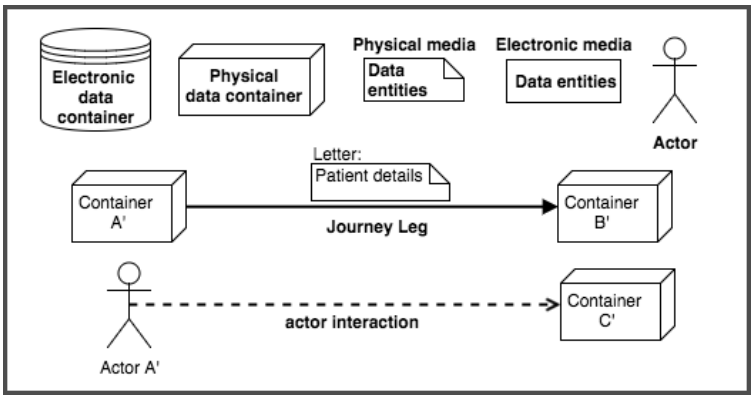


Figure 1: The notation of the data journey model.

the data journey that must be supported for the blood tests results to reach the GP and effective treatment decisions to be made.

The aim of a data journey model is to model the broad movements of a set of data entities through complex networks of people, systems and organisations. Data journey models do not attempt to provide a complete representation of an organisation or its processes; nor are they expected to model the complete set of data movements. Rather, they provide a simplification of reality showing the journeys of the entities of interest within the information infrastructure. Data journey modelling is a lightweight, agile technique that can capture a variety of abstraction levels depending to the needs of the modeller. Depending on the level of abstraction, we have found that data journey models can be completed in as little as two hours.

3.1. Data Journey Model Components

The components of a data journey model (and the notation we propose for them) are given in figure 1. We will now describe each component in turn.

We call the information infrastructure through which the data of interest moves the *data landscape*. It includes both people and technical components that contribute to the creation, storage, transportation and use of information. We distinguish components that store data from those that interact with data (creating it and consuming it).

Data containers are places where data can "rest" (be stored) on its journey through the

data landscape. A container can be in electronic form (e.g., a database, an Excel file, a word document) or in physical form (e.g., file cabinets, desks, pigeon holes). We denote electronic data containers with the database symbol and physical ones with a rectangular box, as shown in figure 1. In the pathology lab example previously introduced, the containers are the GP's desk, the GP reception desk (both storing the request card and blood sample), the pigeon holes of the hospital, the pathology lab secretary's desk, and the pathology system database (storing the blood test results).

Data stored in a container can travel to another container through some already established route. In the data journey model we call these container-to-container movements *journey legs*. A journey leg connects two containers if there is a medium through which they can share data. It allows data to move from a source to a target container to be used for a value-creating step or to await further onward movement. We denote journey legs with a straight line arrow connecting two containers. The direction of the journey leg shows the movement of a set of data entities from the source to the target container, as shown in figure 1.

Each journey leg moves data through a type of *medium*. Media can be of physical or electronic form, and includes paper forms and documents, X-ray films, cassettes, internet messages, emails, etc. For example, the blood test request is moved from the secretary's office to the pathology lab on a piece of card, using a courier service. The test results move from the lab database to the GP's system through an internet connection.

The other types of component modelled in the data landscape are *actors*. These are the people or IT systems that interact with containers to create, consume or transform the data stored in them. Actors are denoted using the actor symbol of the UML notation [25], and their interactions with the containers are shown as a dotted arrow beginning at the actor and ending with the container with which the actor interacts.

Taken together, the legs between the components show the data journeys that are currently supported by the data landscape that is modelled. A *data journey* is a sequence of legs that makes a connected path through the data landscape for a set of data entities of interest, from a point of origin (a point at which the data first enters into the landscape) to a final

destination (a point at which the data is used to deliver some value for some stakeholder). The way-points on the journey may themselves be producers or consumers of data, or they may merely hold data. Journeys may be simple (traversing just one or two legs within a single department) or complex (covering a network of cooperating organisations across wide geographical or organisational distances). They may describe current movements happening in an existing infrastructure, or they may describe planned movements for a proposed new IS development (in which case, certain legs may be missing in the existing data landscape, and need to be added to allow the proposed data journey to take place).

Figure 2 shows the meta model of the data journey (expressed as a UML class diagram) describing the relationships between the elements of the model. In brief, this model states the following. A data journey is a set of one or more consecutive journey legs. A journey leg moves one or more data entities from a source to a target container through either electronic or physical media. A container can hold more than one data entity at any time. An actor interacts with the data stored in a container to consume, transform or create new data entities. Several actors can interact with one container, and the same actor can interact with multiple containers.

3.2. Creating Data Journey Models

The first step in creating a data journey model is to identify the scope of the movement we want to model: i.e., the set of data entities needed at the new development for a value-creating step. For example, a new IT system might be needed that will use data already existing in another system, or a new guideline might require the sharing of data with an external organisation for the first time. Having defined our scope and identified the data entities of interest, we model the containers and actors in the data landscape, and the legs along which the data entities of interest move.

One way of constructing a data journey diagram is using the bottom-up approach: start by identifying individual elements of the journey and then link them together to create the model. Another way to construct a data journey model is the top-down approach. In this approach we first identify the legs of the journey, i.e. from the business processes. We can

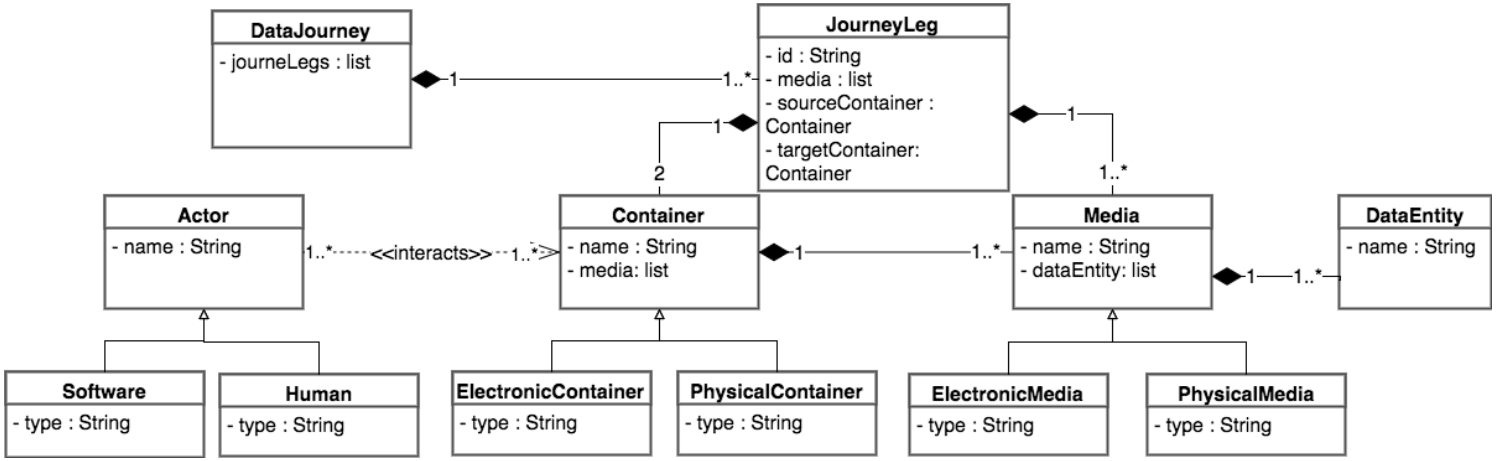


Figure 2: The data journey meta model

then derive the individual elements of the journey from each leg. In either case, the process is iterative; it may take more than one cycle (adding and removing elements) to design the full model. Each cycle refines the model to better represent our interpretation of the data landscape and the data journeys that it can support.

We illustrate the creation of a data journey model for the Pathology Lab example mentioned earlier, using the bottom-up approach.

1. *Identify* the data entities of interest, i.e., the data required at the new development. Also, identify the landscape in which data moves, from the point of entry to the landscape to the point of use.

2. *Identify* the containers at which the data of interest first enter the landscape, those where they are consumed to generate value, and those where they reside *en route* to their eventual consumption. *Add* the identified containers to the model.

3. *Identify* the routes by which data entities move between containers. *Add* the routes to the model to show the possible journey legs. The direction of the arrows denotes the movement of data from a source to a target container. *Number* the journey legs for future reference.

4. For each journey leg, *add* to the model the medium by which the data entities are moved across the leg from sourve to target. We distinguish between physical and electronic types of media. Within each media symbol, *note* the data entities that move along it.

5. *Identify* the actors interacting with each container to either create, consume, or transform the data entities of interest. Numerous actors can interact with the data entities we identified. However, we focus on finding the ones who interact with data to produce some value within the scope of the journey. *Add* to the model the actors using the actor symbol. *Label* each actor with its role in either creating, consuming, or transforming the data, and its position (in the case of systems state its name).

6. *Connect* each actor with the container it interacts with, using a dashed line arrow beginning from the actor and ending at the container. *Label* each interaction with the process/action achieved.

The results of each of these steps, applied to the simple Pathology Lab example, are shown in figures 3 and 4. At the end of step 6, the data journey model of the existing data landscape is complete.

4. Predicting Socio-Technical Costs/Risks in Data Journey Models

In this section, we propose a method that uses the data journey model described in the previous section, to predict places of high cost and risk in a planned new software development. The emphasis on a lightweight, low-cost method is based on the limited time, and often, budget of managers and employees of large organisations to invest in deciding whether to proceed with a new development or not. Hence, we must focus on obtaining just the bare minimum of information needed, and ideally only on information that is readily available or cheap to acquire. Specifically, we need a method that guides us in:

- Modelling the necessary parts of the existing information infrastructure of an organisation, but also the parts of the planned new IT development to be integrated in the infrastructure.
- Modelling the movement of data from a point of entry in the existing infrastructure to the point of use by the new consumer.
- Predicting places of the journey that, because of some socio-technical barrier, can impose high costs and risks on the new development.

Our method begins by identifying the set of existing data entities that must travel to the new IT development. We then model the information infrastructure in which data entities currently exist and the journeys already happening within it, using the data journey model. Next, we add to the model details of the new development, in the form of the new data

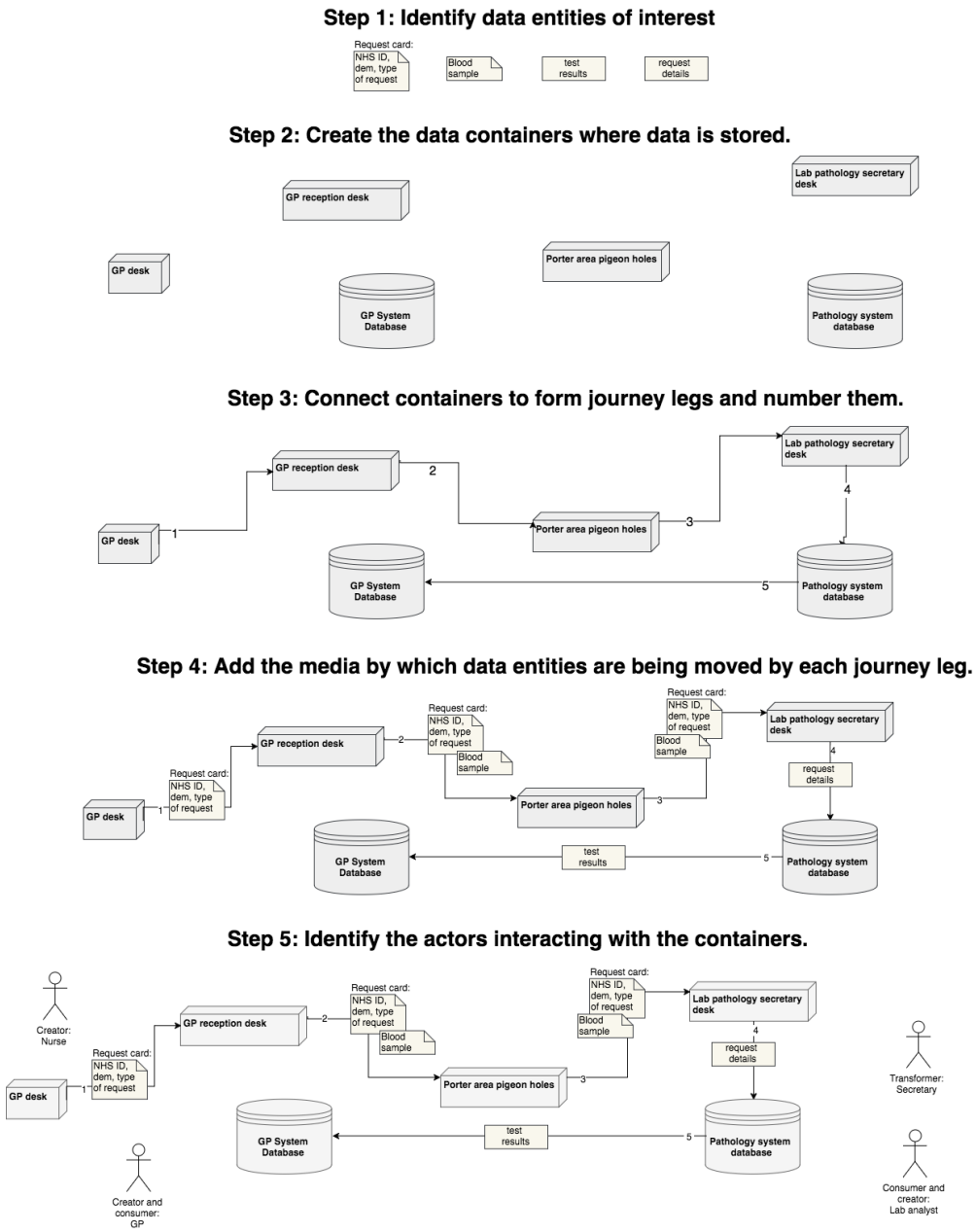


Figure 3: An example illustrating the steps to construct a data journey model (continues in figure 4)

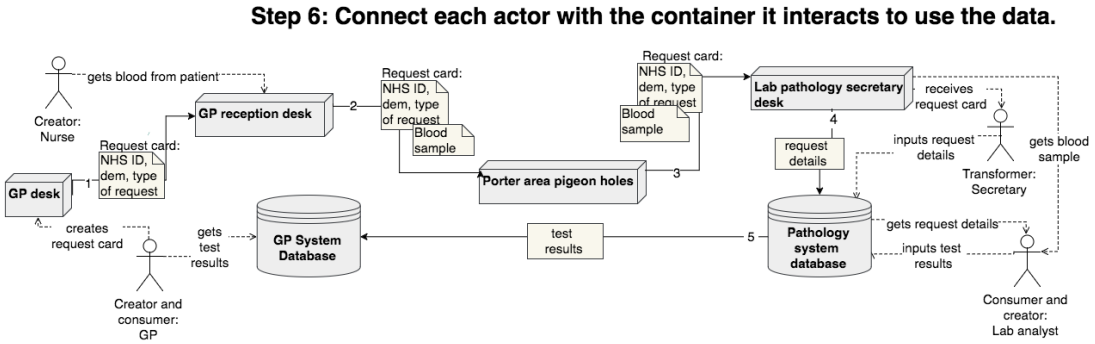


Figure 4: An example illustrating the steps to construct a data journey model (continued from figure 3)

journeys needed to support it. Once we have modelled both the existing and new data journeys, we overlay onto the model socio-technical information to help us identify places in the journeys that can cause high costs or are subject to risk. Because our model includes movement of data between people, and not just between computer systems, it has the ability to identify a wide range of costs and risks, including the people-oriented issues that caused the greatest problems in the NHS case studies from which we derived the approach.

Figure 5 summarises the four steps of the method. We have already described step A, the creation of the data journey model for the existing information infrastructure. In the remainder of this section, we explain steps B to D, illustrating each one as we go with the resulting models from the Pathology Lab example.

4.1. Adding The Planned Data Journeys

Having modelled the existing landscape in which the data entities of interest move, we add to the model the planned new IT development and the new journeys that are needed to move data from the existing landscape to the planned new consumer. This may involve adding new containers, actors and legs to the model.

For example, suppose that an external government agency requires NHS pathology labs to share demographics data on the tests they do, as part of an initiative to make workload sharing between local labs more effective. The particular lab we are supporting needs to work out what new software systems and business processes are needed to meet this statutory

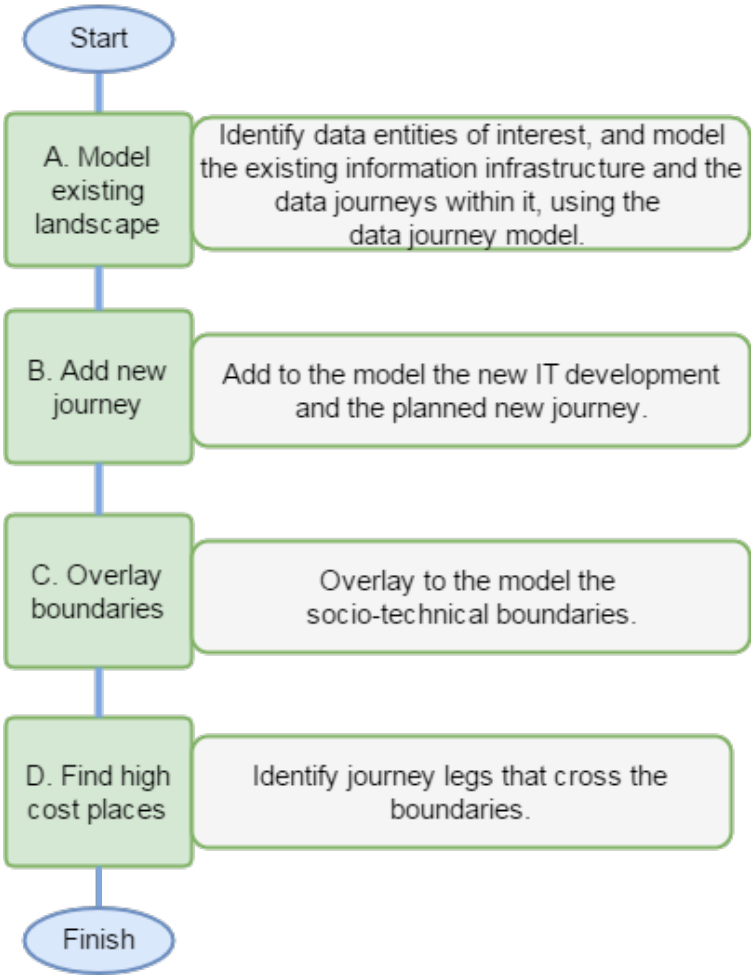


Figure 5: Overview of the cost/risk prediction method

requirement. We add to the model the new development and journey as follows:

1. *Add* to the landscape any new containers required by the new development to store the required data. In the case of the Pathology Lab example, the only new container is the database hosted by the external agency, into which the demographics data must be fed.

2. *Identify* the existing containers that store the entities of interest, and *select* the one which will supply the data for the new development. *Connect* this container with the most appropriate target container to complete the required data journey. *Annotate* the planned legs with information about the media through which data will be moved.

3. *Add* any new actors who will interact with existing or new containers to create some value relating to the task in hand. In this case, only one new actor is required: the staff at the external agency who will consume the data to create their reports.

Having followed above steps we now have the data journey model of both the existing landscape in which data of interest moves, and the new IT system and journey to be implemented. The complete data journey model of the pathology lab example is given in step 7 in figure 6.

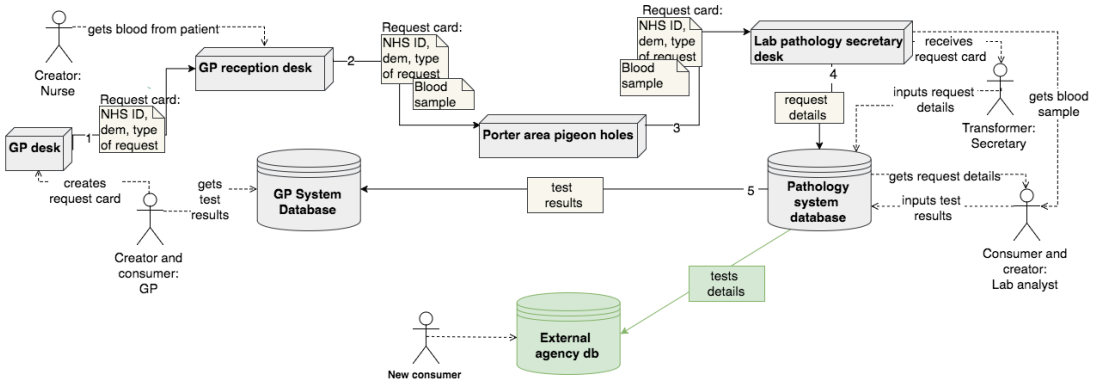
4.2. Overlaying Socio-Technical Risk Factors

Having created a model of the existing and planned data journeys, the next step of our method is to find the places in the journey that are costly/risky. The analysis of the case studies described in section 2 showed that costs and risks arise when data is moved between two organisational elements which differ in some socio-technical way that is significant to the interpretation of the data. These are the places where the portability of the data (i.e., its ability to retain its meaning when moved to a context other than the one it was originally designed for) will be put under stress, where errors can occur when the differences are not recognised, and where effort must be put in to resolve the differences. Some of the most common socio-technical discrepancies we found in the NHS case studies we analysed are:

- When data moves from a source container to a target container of a different media (i.e., physical to electronic), then a transformation cost exists, either before or after the transportation of the data, that can lead to decreased data quality at the target side.

- When data moves from a source container to a target that belongs in a different context (i.e., organisation, geographical area, culture, etc.), then a bridging cost is imposed on one or both sides of the flow. For example, in the case of organisational discontinuity (e.g., when data needs to move to an external agency) there is the bridging cost of complying with the data sharing agreements, governance, and ethical guidelines needed to export sensitive data outside the organisation that created it.

Step 7: Add the new system and required journey.



Step 8: Form boundaries by clustering similar entities together.
and
Step 9: Identify journey legs crossing the boundaries.

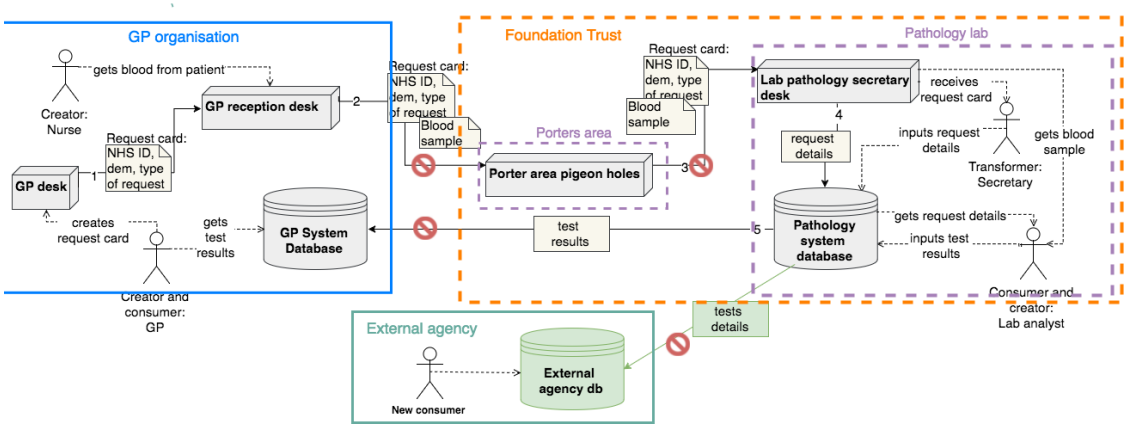


Figure 6: Predicting costs and risks for the Pathology Lab data journey model

- When data moves from a consumer to a producer, a difference in a property of either the source or the target introduces a transformation cost to the movement. For example, if data moves between people with different levels of expertise and vocabularies, then a risk of *clash of grammars*[1] may cause the data to be interpreted differently in each side of the movement. This can result in loss of data quality at the target side.

What must we add to the data journey model to allow us to identify points where discrepancies of the kind noted above might lie in wait? Any such information must be cheap to acquire, since there is little value in predictions that cost a significant fraction of the actual development costs to produce. For some of the discrepancies we identified, the information is readily available. For example, it is normally well known to stakeholders when information is stored on paper, in a filing cabinet, or in electronic form. However, other properties, like cultural and communication differences between staff are less obvious. For these properties we use a proxy; some piece of information which is cheap to acquire, and approximates the same relationship between the actors and containers as the original factor. For example, we use salary bands as a proxy indicator for the difference in actors' vocabularies, on the grounds that a large difference in salary bands between actors can indicate a different degree of technical specialism.

We use this information to form *boundaries* in the data journey model. A boundary is a dividing line that clusters data journey elements with similar properties together. We overlay boundaries on top of the data journey model to identify the places where data is moved between containers with discrepancies that can potentially introduce costs and risks to the journey. We use the following rules and proxies for indicating the presence of a boundary:

Change of media: the medium of the source container of a journey leg is different from the medium of the target container.

[1]That is, the meaning of the data being altered because of a cultural, experience, knowledge or other difference between the actors [34].

Organisational discontinuity: the source container of a journey leg belongs to a different organisational unit from the target container.

Actors properties: the actor creating the data at the source container of a journey leg has a different salary band from the actor consuming it at the target container.

Other, more costly, types of boundary are also likely to exist [35]. Here, we use the ones that we found to be most prevalent in the NHS case studies we examined and cheap-to-acquire in practice.

To identify the presence of boundaries in a data journey model, we group together the elements of the data journey diagram with similar properties based on the identified rules. For example, we group together all the data journey elements belonging to the same organisation. We then overlay the groupings on top of the data journey model to form boundaries. Step 8 in figure 6 shows the organisational boundary for the Pathology Lab example. The data journey elements belonging to the GP organisation are shown within the blue boundary, whereas the ones belonging to the hospital are within the orange dotted lines. Within the hospital boundary we have the two internal boundaries: the containers belonging to the hospital porters' area and the elements belonging to the hospital's pathology lab.

4.3. Identify Points of Predicted Cost/Risk

The final step of our method is to identify the places in the data journey model where high costs (or the risk of high costs) are predicted. To find those places, we identify the journey legs that move data across a boundary. If the source of a journey leg belongs to a different side of a boundary than the target, then a key discrepancy between data journey elements exists, and a prediction of high cost/risk is made. In step 9 of figure 6 the costly journey legs are shown with a red warning sign. For example, based on the organisational boundaries, we identified 3 predicted cost/risk points. These are the places in the journey where data moves away from the context where it is in established use. Table 1 gives the likely costs and risks that a boundary might impose to the development if a journey leg crosses it.

Boundary crossed	Likely costs and risks
Organisation	Sharing data outside the immediate organisational unit can result in a number of administrative costs, such as reaching and complying with data sharing agreements, as well as complying with wider information governance and ethical requirements.
Change of media	Entry of data on paper into an electronic target system is a time consuming process, typically done by administrative staff who may not have a strong understanding of the meaning of the data they are entering. Errors can be injected that may significantly reduce the quality of the information stored at the target side.
Actor role	There is a risk of "clash of grammars", and a cost of lower data quality at the target side. For example, data entered into a system by secretarial staff can contain errors if, to be interpreted correctly, the information requires medical knowledge/vocabulary that the secretarial staff lack.

Table 1: Costs/risks imposed when a boundary is crossed

All three types of boundary should be overlaid onto the data journey model to create a cost/risk heat map. Legs that cross multiple boundaries should be regarded as the points of highest predicted cost/risk.

5. Evaluation in an NHS Hospital

To determine whether this method could predict points of cost and risk in real systems, we conducted a retrospective evaluation through a case study in health care. In the case study, we examined a data movement example in a radiology department of an NHS Foundation Trust in the UK. Prior to our study, hospital staff had identified costs and delays in their

old IT system handling the appointments and patient data, and had recently introduced a new improved system. To evaluate our prediction method, we looked back at the old system before hospital staff made any improvement efforts to reduce delays and costs. We applied our method to the old system, without knowing what improvements had been made, to predict high cost places in the data journeys. Only then did we model the new system, to compare our predictions with the changes made by hospital staff.

More specifically, to retrospectively evaluate the cost predictions of our proposed method we:

- Conducted semi-structured interviews with two NHS domain experts working at the Informatics department of the Trust to gather domain knowledge, such as data movements, organisational structures, available information systems, data formats and staff roles.
- Modelled the data journey of the *old* system before improvements were made by hospital staff.
- Applied our method on the *old* model to predict journey legs of likely high costs and risks.
- Modelled the data journey of the *new* system produced as a result of the hospital's own improvement efforts.
- Identified the journey legs that hospital staff improved by comparing the two models ('old data journey' and 'new data journey').
- Compared the predicted costly journey legs of the old model with the staff's improvements to assess the accuracy of our predictions, with the help of the NHS domain experts.

This section describes the approach we took, and presents the models and predictions created, alongside the assessment of the results.

5.1. Success Criteria

Before embarking on the evaluation, we defined a set of success criteria to assess the outcomes of the evaluation. We didn't expect our model to predict all the changes made by the domain experts in the new system, since that would require detailed and complete modelling, and a greater investment in time from domain experts. Instead, we evaluated whether our lightweight model can cheaply and quickly predict the major points of high cost/risk, where savings can be made, while investing only moderate resources (time, effort, money) in the prediction. We defined our goals for assessing the lightweight property of the method and the accuracy of the predictions as follows:

Domain Expert's Time We kept track of the time invested by hospital staff to provide us with domain information needed to make the models. We set the threshold to be 1 working day (up to 7 hours) of interview/modelling time for each 'old' and 'new' model, since hospital staff's time is valuable.

Modelling Time We also monitored the person-hours required to create the 'old' model and predict places of high cost and risk. Building the model and predicting costs must take a small fraction of the system development time, so based on the scale of this case study we set the threshold to 1 working day (up to 7 hours) for this, too.

Prediction accuracy We set a success threshold for the method of doing as well as the hospital staff in predicting points of cost and risk. That is, we would regard the method as successful if it could perform *at least as well* as the hospital staff in predicting points of cost and risk. The new processes and IT systems had been in place at the hospital for some time when the evaluation was carried out, and the hospital staff had seen evidence that the benefits they hoped for from them had been realised. We therefore took the pragmatic decision to assume that the hospital staff had correctly identified a set of cost improvements in the old system. However, we also assumed that there might have been other potential cost saving measures that were not implemented by staff (perhaps because of limited resources, or because the opportunity was not identified at the time).

> Since our model aims for a good-enough answer quickly, we do not expect it to be accurate and complete. However, every inaccurate prediction has the cost of further investigating it. Therefore, we set the conservative goal of considering the model to be accurate if it produces fewer wrong predictions than correct predictions.

As with any large commercial or governmental institution, some aspects of the details of the case study are confidential. Although the models used to evaluate our method were based on the actual case study, here we present a generalisation of the model, typical of what might be used in a range of NHS hospitals. The results of the evaluation presented are the original ones produced from the actual case study.

5.2. Modelling the Old Hospital System

In the case study we examined the movement of data that occurs when a GP needs to decide on an action plan for a patient who may have a fractured bone. In the original version of the system (called the *old system* in what follows), the GP requested an X-ray to be taken at the local hospital's radiology department by filling in a request card. The request card was then sent to the hospital's radiology department by courier. An appointment was made for the patient to attend the radiology clinic. When the patient arrived, a radiographer took an X-ray image of the injured area. A radiologist reviewed the image and dictated a report with the X-ray findings. Then, the secretary transcribed the report into the system, printed it and posted it to the GP to decide on an action plan.

We conducted semi-structured interviews with two domain experts working in the Informatics department of the hospital, to gather information needed to create the data journey model. We needed to know the data entities needed by a GP to decide on treatment: these are the patient's identification details and the report containing the X-ray findings. We also needed information about the containers, actors and movements of data that took place in the old system. The early versions of the data journey model was created off-line, without the domain experts being present, in order to reduce the possibility (as far as possible) that the experts might bias its contents through knowledge of the points of cost and risk. The domain experts then validated the model. As an extra check, the data journey model was

validated independently by a third member of hospital staff, the PACS (Picture Archiving and Communications System) Imaging Informatics Manager, to ensure that the movements shown were representative of the hospital's processes.

Appendix A on page 41 gives an example of data movement processes in a typical NHS FT, similar to the subject of our case study. With the help of the domain experts, we designed the data journey model given in figure 7. An example of the process we followed to create the data journey model is given in Appendix B on page 42.

5.3. Predicting Points of High Cost/Risk

Once we have modelled the journey of the data, we overlaid the socio-technical boundaries onto it, to predict places in the journey of likely cost and risk. We overlaid the three types of boundaries of our method: the organisational, change of media and actors role boundary. The data journey models with each of the overlaid boundaries are given in Appendix C.

Based on our method, the journey legs that cross a boundary are the ones predicted to introduce costs and/or risks. Figure 8 overlays all three boundaries to form a *boundary heat map*, and highlights the journey legs of interest. Journey legs with a bold red arrow cross one type of boundary, whereas legs with a bold red double arrow cross two types of boundary. We found no leg that crossed all three boundaries in this study. Table 2 lists all the journey legs that crossed one or more boundaries, and describes likely costs and risks the crossed boundaries might impose on the movement.

Journey Leg	Crossed boundary	Likely costs and risks
2	Organisational, and Actors role	Data moves away from the GP organisational unit to the Foundation Trust indicating the cost of complying with data sharing agreements, governance and ethical guidelines. Also, data created by the GP are used by another actor of different position/role, implying a risk of clash of grammars and lower quality of data moved to the target.

3	Change of media	Data moves from the physical container of the clinical reception desk to the electronic container of the radiology's system database, causing a data entry cost. There is also the risk of injecting errors that can reduce the quality of the information at the target side.
4	Organisational, and Actors role	Data moves between two organisations, the FT and the community, indicating the cost of complying with data sharing agreements, governance and ethical guidelines. Also, data created by the secretary of the FT are used by a user of different role, the patient, risking clash of grammars and lower data quality.
5	Organisational	Data moves from the archives department of the FT to the radiology department indicating the cost of complying with data sharing agreements, governance and ethical guidelines.
8	Actors role	Data created by the radiographer are used by another actor, the radiologist. The risk of clash of grammars exists that can cause lower data quality to the radiologist.
9	Actors role	Data created by the radiologist are used by another actor, the secretary. The risk of clash of grammars exists that can cause lower data quality to the secretary.
10	Change of media	Data moves from the physical container of radiology secretary's desk to the electronic container of the radiology's system database causing a data entry cost. There is also the risk of injecting errors that can reduce the quality of the information at the target side.
11	Change of media	Data moves from the electronic container of the radiology's system database to the physical container of the radiology secretary's desk. There is the cost of printing and transferring the data to the target side.
12	Organisational	Data moves from the radiology department of the FT to the archives department indicating the cost of complying with data sharing agreements, governance and ethical guidelines.
13	Organisational	Data moves from the radiology department of the FT to the GP organisation indicating the cost of complying with data sharing agreements, governance and ethical guidelines.

14	Change of media	Data moves from the physical container of the GP reception desk to the electronic container of the GP's system database causing a data entry cost. There is also the risk of injecting errors that can reduce the quality of the information at the target side.
15	Change of media	Data moves by the electronic container of the GP's system to the physical container of the secretary's desk. There is the cost of printing and transferring the data to the target side.
17	Actors role	Data created by the GP's system are used by an actor of different role, the GP. There is a risk of clash of grammars that can cause lower data quality to the radiologist.

Table 2: Costs and risks predicted by the method.

5.4. Modelling the New Improved Hospital System

In the old system, X-rays were captured on large X-ray film and stored in the Archives room. This caused delays in transporting the films to the place of need, while investing time and resources to file, capture, and transfer the films and the patient's notes. Also, there was a small degree of risk of losing the films. In the new system, the old X-ray machinery was replaced with state-of-the-art electronic equipment that captures and stores X-ray images in digital form. X-ray images are now uploaded to the PACS system, and are no longer printed on X-ray films. The new digital images can be easily modified to highlight and magnify the area of interest, and transferred around the hospital as needed.

The PACS system is integrated with the Computerised Radiology Information System (CRIS) responsible for receiving referrals, booking appointments, and managing patients. CRIS replaced the old radiology system and is fully integrated with key hospital information systems such as the Patient Administration System (PAS), the Order Communications, and the Electronic Patient Records system (EPR). The data journey model of the new system is given in figure 9 on page 30.

To evaluate the predictions of our method when applied on the old data journey model, we had to identify the improvements hospital staff had made to the journeys of the data. To do so, we compared each journey leg of the old model with its closest corresponding

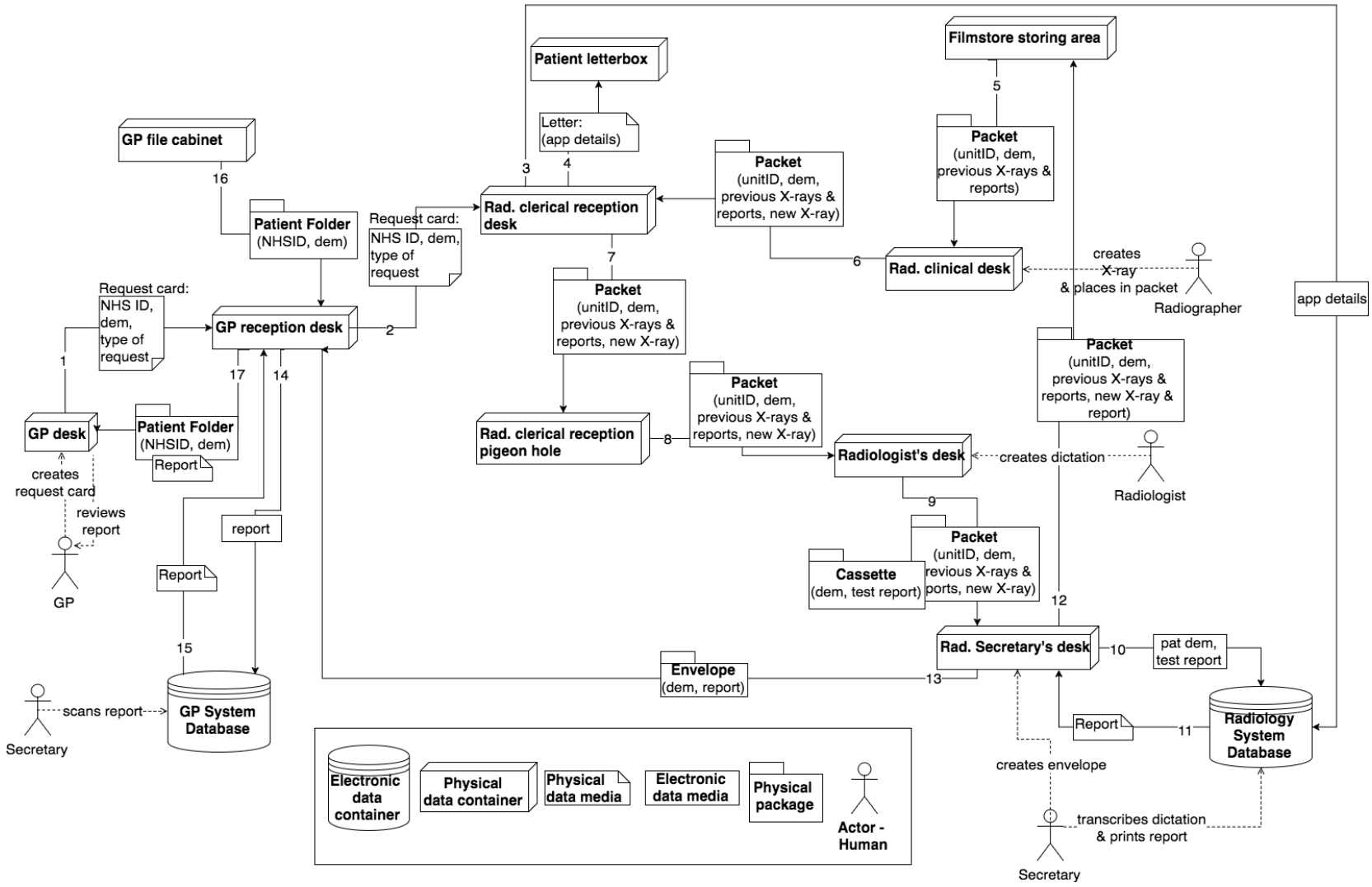


Figure 7: Data Journey Model of the 'Old' Radiography System

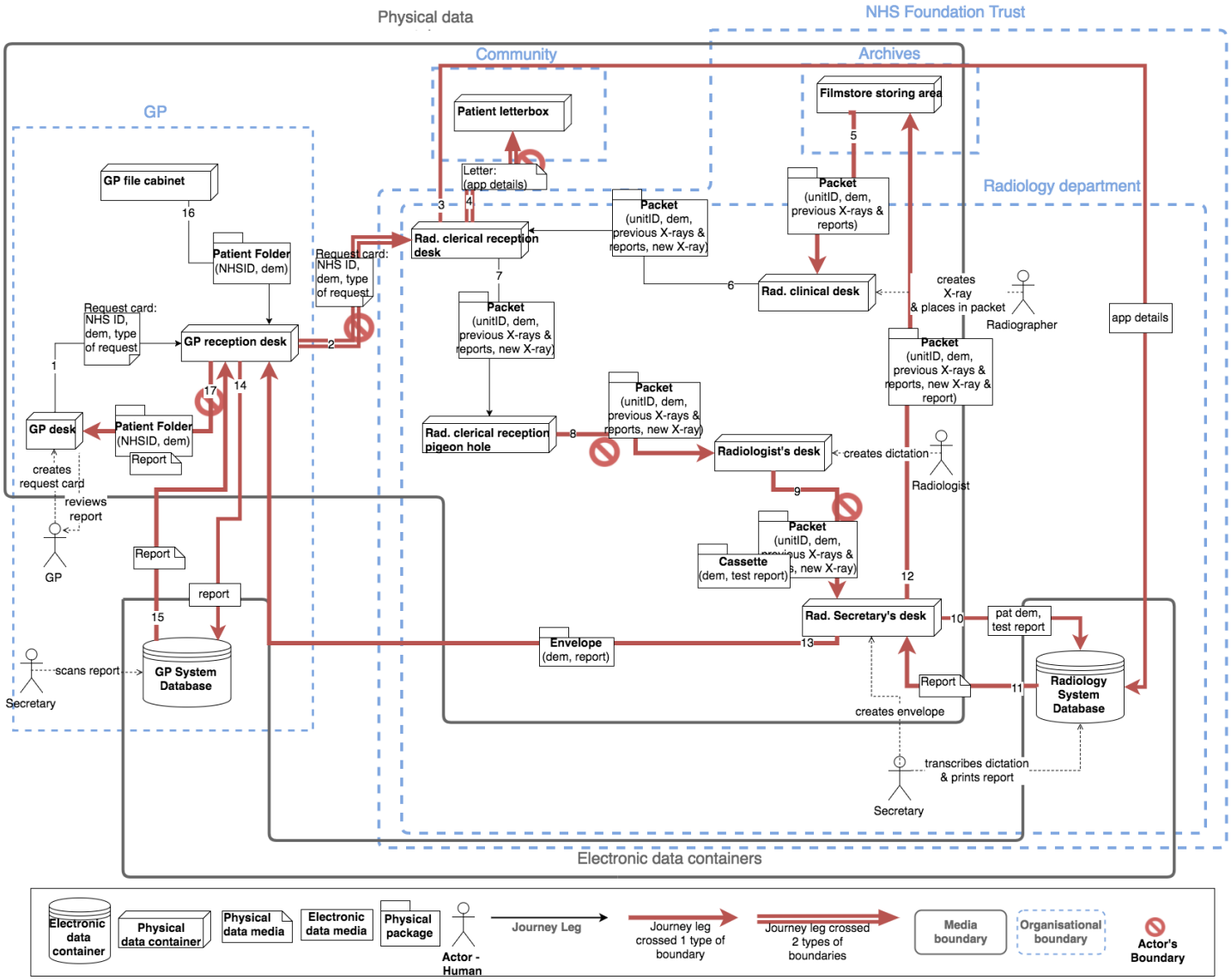


Figure 8: Boundary Heat Map Overlaid onto the Model of the Old System

leg in the new data journey model, to find whether the predicted costly movements have been removed or replaced in the new system. Table 3 shows the differences between the two models, highlighting the changes made to the movements of the old system and the corresponding new journey legs.

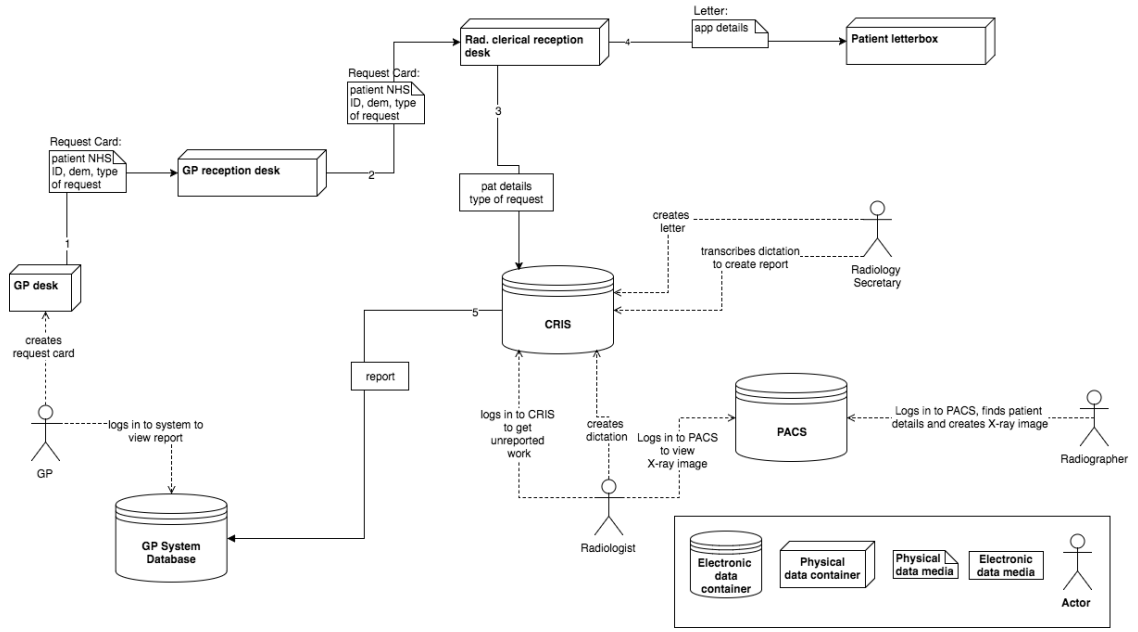


Figure 9: Data Journey Model of the 'New' Radiography System

5.5. Evaluation Results and Discussion

In this section we evaluate the accuracy of our method at predicting places of cost and risk using the data journey model and boundaries approach. By comparing the old data journey model with the new model, we identified the set of journey legs that hospital staff assessed as costly and replaced in the new model to reduce the overall costs of the system.

Given the nature of this project, we assess a prediction as accurate-enough if the predicted journey leg has been also found to be costly by hospital staff, and changed in the new model. While we assume that hospital staff correctly identified some cost saving opportunities in the old system, we also consider that other opportunities for potential cost savings might exist that were not implemented in the new system. To evaluate the accuracy and feasibility of the rest of our predictions we asked the two NHS domain experts to independently review them.

Old leg	Changes made by hospital staff	New leg
1	No change.	1
2	No change.	2
3	Different target container, the radiology system is replaced by CRIS system.	3
4	No change.	4
5	Leg removed by replacing physical packet with electronic data saved in PACS.	-
6	Leg removed by replacing physical packet with electronic data saved in PACS. Radiographer creates electronic X-ray image in PACS.	-
7	Leg removed by replacing physical packet with electronic data saved in PACS.	-
8	Leg removed by replacing physical packet with electronic data saved in PACS.	-
9	Leg removed by replacing physical cassette with electronic data saved in CRIS. Radiologist accesses patient details through PACS, and dictates report in CRIS.	-
10	Leg removed by replacing physical cassette with electronic data saved in CRIS. Secretary accesses dictation and transcribes report in CRIS.	-
11	Leg removed. The report is not printed, as will be electronically sent to the GP.	-
12	Leg removed by replacing physical packet with electronic data saved in PACS.	-
13	Leg replaced by using electronic report sent directly to the GP system.	5
14	Leg replaced by using electronic report sent directly to the GP system.	5
15	Leg removed. GP accesses report directly from the GP system.	-
16	Leg removed. Archives still exist, but not used in everyday processes.	-
17	Leg removed. GP accesses report directly from the GP system.	-

Table 3: Changes made to the old system by hospital staff.

We asked the experts to assess a prediction as 'valid' if they could foresee a straightforward mechanism for eliminating the predicted leg which would result in a cost saving. We also asked them to comment on the costs and risks they faced while working with the old system.

Table 4 gives the assessment of our method's predictions. It shows whether the costly predicted journey legs were removed or replaced in the new model (i.e., meet our main accuracy criterion). For each prediction, the table presents the views of the NHS domain experts on the costs and risks imposed by the respective boundary, along with an assessment of the validity of each prediction. Finally, we give a final assessment of the accuracy of each prediction. We categorise the final assessment as follow:

True Positive (TP) A prediction is assessed as a TP if the predicted cost was either removed in the new model *or* assessed as 'valid' by the domain experts.

False Positive (FP) A prediction is assessed as a FP if the predicted cost wasn't removed in the new model *and* the domain experts assessed it as a 'not valid' prediction that wouldn't lead to cost savings.

False Negative (FN) A journey leg is assessed as a FN if it was not included in our predictions but the domain experts foresaw likely costs and risks associated with it.

True Negative (TN) A journey leg is assessed as a TN if it wasn't included in our predictions *and* it wasn't changed in the new model *and* the domain experts didn't foresee any likely costs and risks associated with it.

Old journey leg	Crossed boundary	Journey leg removed in new model?	Domain experts assessment	NHS domain experts view on costs and risks	Final assessment
1	None	-	-	-	TN
2	Organisational	No	Valid	There is a transportation and postage cost transferring the request card from the GP to the radiology department.	TP

2	Actors role	No	Valid	Since the request card is filled in by a different person than the one using it, there might be a missing information cost. In the case of an uncompleted request card (happens frequently), the process is disrupted. Costs include time to call the GP chasing missing information, time and effort to complete another request card, resources of the replacement card. Also, these extra costs imposed on the process can presumably damage the reputation of the GP and FT.	TP
3	Change of media	No	Valid	Inputting data into the system requires time and effort of clerical staff. Also, it introduces the risk of injecting errors. The risk of injecting errors into the system is higher since data was created by a user other than the hospital clerical staff (i.e. the GP secretary). Also, there is duplication cost since data already existing on the request card are duplicated into the system.	TP
4	Organisational	No	Valid	There is a postage cost imposed when sending the letter to the patient. Resources are needed like paper, stamps, envelopes and the postal franking machine.	TP
4	Actors role	No	Valid	Time and effort of the clerical staff to print and prepare the letter. Also, since the appointment date and time are selected by the radiology secretary but actually used by the patient, there is the risk of the unavailability of the patient and the cost related with the patient cancelling or rescheduling the appointment.	TP

5	Organisational	Yes	Valid	There are searching and transferring costs. Time and effort of clerical staff to find the patient's packet from the archives and transferring it to the radiology clinical area. Also, resources are needed to transfer the big packets across units, mostly by using a trolley.	TP
6	None	-	-	-	TN
7	None	-	-	-	TN
8	Actors role	No	Not valid	The radiographer creates the X-ray image which is then used by the radiologist. Mistakes made because of the different actors are possible, but unlikely.	FP
9	Actors role	No	Valid	The dictation is created by the radiologist but transcribed into the system by the secretary. This introduces the cost of mistakes inputted into the system. This is actually a common phenomenon since secretaries do not share the same knowledge and experience of radiologists to always fully comprehend what they meant to say in the dictation.	TP
10	Change of media	Yes	Valid	There is the cost of the time and effort of the secretary typing data into the system and the risk of injecting mistakes and errors.	TP
11	Change of media	Yes	Valid	There is the stationery, time and effort cost of the secretary printing the report.	TP
12	Organisational	Yes	Valid	There is a transportation cost transferring the packet with the patient's information back to the archives. Time and effort of clerical staff are needed. Also, resources like the trolley are required.	TP

13	Organisational	No	Valid	There is a transportation cost transferring the enve lope containing the report to the porter's area of the FT, staff time and effort.	TP
14	Change of media	Yes	Valid	There is the cost of the time and effort needed by the GP secretary to scan the letter into the GP system. Also, there is the risk of duplicating information already existing at the FT.	TP
15	Change of media	Yes	Valid	There are printing costs.	TP
16	None	-	-	Cost of retrieving and filing patient records.	FN
17	Actors role	No	Not valid	The report is created by the FT radiologist which is now used by the GP. Often, there is the cost of a follow up appointment with the patient. However, this is not caused by the actors barrier.	FP

Table 4: Assessment of the Accuracy of the Predictions

From the table above, we see that the old data journey model has 17 journey legs. Of these, 13 cross a boundary, and hence are included in our predicted points of cost/risk. Some of the 13 legs crossed more than one boundary. In the above table, the two journey legs crossing more than one boundary are presented as different predictions (since the likely cost imposed by the different boundaries can be of another type) making the total number of predictions 19.

More than half of our predictions (13 out of 19) are assessed as true positives. Of these, 6 were removed in the new model, while the remaining indicate boundaries that still exist in the current system and were assessed by the NHS domain experts as 'valid' predictions.

Of the 19 predictions, 2 were assessed as false positives, since they weren't removed in the new model and the domain experts assessed them as 'not valid'. Although our method predicted the existence of boundaries at these points, the costs that the domain experts have identified for the respective journey legs didn't stem from the predicted boundaries.

	True	False
Positive	68%	11%
Negative	16%	5%

Table 5: Contingency Table Showing Evaluation Results

Both journey legs crossed the actor boundary: data moved from the radiographer to the radiologist and from the radiologist to the GP. Although these actors have different salary scales, they share the knowledge and experiences needed to comprehend information created among them. This indicates a limitation of the salary band proxy used to predict the cost of clash of grammars between actors and further research is indicated.

Of the remaining predictions, 4 did not cross any boundary. The domain experts couldn't identify any costs or risks related to 3 of these 4 legs, making them false positives. However, the domain experts did foresee additional costs and risks in the case of the remaining leg (leg 16). This was assessed as a false negative since it crossed no boundary in the model but domain experts identified a cost of retrieving, filing and transferring the patient folder. Our method didn't predict the cost of transferring physical data, since it only captures costs of transforming data from physical to electronic format and *vice versa*. Other transfer costs in the model (e.g., journey legs 2 and 5) were captured by the organisational boundary, as different organisational units tend to be in different geographical areas, suggesting further research is needed on the intersection of the organisational and transportation costs.

Table 5 shows the ratio of true positives, false positives, true negatives and false negatives in our predictions.

In addition, we asked the domain experts to comment on which leg of the model they judged to be the most costly/risky and that if eliminated would have the biggest long term benefit to the organisation. Both domain experts *independently* assessed prediction number 2 (actors boundary on journey leg 2) to be the most costly. This is one of the two journey legs (numbers 2 and 4, where data are transferred from the GP to the FT and from the FT to the patient's home, respectively) that our method predicted to have the most barriers and hence be most costly, as illustrated in the heat map (figure 8). The prediction of the

journey leg 4 was also ranked highly by the experts (4th and 5th most costly leg).

Our method identified a majority of the journey legs of the old model (14 out of the 17) as involving physical media. Physical legs have significant stationery, printing and transportation costs as well as the risk of data quality loss (duplication, timeliness, incompleteness, consistency, etc). This is an issue identified by the hospital staff and was reduced to only 3 physical legs in the new model, as can be seen in the new journey model (figure 9).

Based on the comments of the domain experts, another significant cost is imposed on the organisation when journey legs cross organisational boundaries. Five journey legs of the old model crossed organisational boundaries, of which only two didn't have any governance issues (numbers 5 and 12), since they move data within the same hospital unit. According to the experts, the legs with no governance barriers were the ones they targeted for elimination. This was because governance is one of the hardest and most complex barriers to remove. Based on the experts' opinions, organisational barriers (data moving across organisational boundary) are the hardest to resolve because of the myriad of governance regulations that organisations need to comply with, especially when the data moved are sensitive patient information. Change is easier and quicker within an organisational unit than across organisations. Also, the three journey legs in the GP organisation of the old model weren't changed in the new model, for the obvious reason that the changes were driven by hospital staff in the radiology department, who had no jurisdiction to enforce change at the GP organisation.

Finally, we interviewed the NHS domain experts on the types of cost and risk that can affect an organisation. Their comments reflect on the findings of our method:

> *"Costs are indicative of both the process and the actors involved in the process. For instance, costs would be higher in respect of processes involving higher skilled personnel, such as time taken by GPs to complete request cards at their practice, and the actual X-Rays performed by the Radiographers. Similarly, from an administrative perspective, costs would be higher where either duplication or other errors are introduced (due to miscommunication, misunderstandings, etc) by medical secretaries. At the lower end of the cost spectrum would be*

the lesser skilled personnel such as clerical staff and porter staff to transport data/information from different end-points such as from the GP practice to the Secondary care facility (i.e., Acute hospital).” - NHS Domain Expert #1

“The majority of the total cost is the human resources. For example the time the GP needs to fill in the request card, and so on. Another major cost is the acquisition and maintenance of clinical and clerical equipment, like the X-ray machine, cassettes, etc. The least expensive costs to the organisation are the stationery and printing costs.” - NHS Domain Expert #2

A property of our method is to predict costly points in a lightweight manner. Although, more information invested in the method will likely provide more accurate results, the time required to acquire this information will be taken away from the domain experts', clinicians', or managers' valuable time. Our aim was to develop a lightweight method that does not demand much time spent on acquiring needed information, designing the models and applying our method. We looked for information that is cheap to acquire, but also points out areas of cost and risk. Because of the lightweight property of our method we do not expect all our predictions to be accurate. However, every inaccurate prediction when reported to potential managers or stakeholders has the cost of further investigating it. Therefore, we consider the model to be accurate if it produces *more* correct predictions than wrong predictions. The results of our evaluation showed that we obtained significantly more correct predictions than false ones.

Apart from the accuracy of the predictions, another criterion of the success of our method was the time needed to collect required information and the effort put to predict barriers. Records of our interviewing times with the NHS domain experts show that they are within our set threshold of 1 working day. We had 3 meetings of an hour each with both experts present (domain expert #1, and #2), and 2 further one-hour meetings with just one of them (expert #2), totalling 8 hours of combined time. However, the records of the time spent modelling turned out not to be useful. This is because we refined the model while working on the case study, and so ended up repeating a significant amount of the modelling work.

We cannot therefore assess the success of our method against this criterion from this study. Later work has suggested that modelling does not add much to the time needed to acquire the domain information, and indeed can be carried out by domain experts themselves after less than one hour of training [35].

Results on both time taken and accuracy of the predictions are within the set thresholds, indicating that our method is lightweight and can well predict places of high costs and risks in the journey of data among organisations. Table 6 summarises our success criteria, the expected and actual times and results.

Success Criteria	Expected outcomes	Actual outcomes
Accuracy	The accuracy of the predictions. At least 50% of the predictions must be TP, while FP predictions must be fewer than TP.	13 out of the 19 predictions were TP (68.42%), while only 2 were FP (10.53%). The rest 21.05% were TN and FN.
Time	The time invested by the staff of the organisation to give us domain information. We set the threshold to 1 working day.	We had 3 meetings of approximately one hour each with NHS domain expert #1, and 5 one-hour meetings with domain expert #2.

Table 6: Evaluation Against Success Criteria

6. Conclusion

The analysis of 18 case studies from the NHS domain showed that although movement of data is vital, it can be affected by several socio-technical factors that can impose high costs on the development of new applications. Given that these costs are often underestimated, we need a way to quickly identify and predict barriers of data movement, ideally before initiating any development.

In this paper we proposed a new low-cost method that uses cheap-to-acquire socio-technical information to predict places of high costs when an existing dataset moves to a new development. The method is based on a lightweight model, called data journey model,

which conceptualises the journey of a set of data from their original location in an information infrastructure to the new development, through a complex network of systems, people, and organisations.

To test our method, we conducted a retrospective evaluation of a real world case study from the NHS domain. We modelled the journey of data in the radiology department of a FT before and after a major information infrastructure redesign. We compared the two models to find if our predicted places of high costs in the old model have been overcome in the new. We also interviewed two NHS domain experts to assess our models, and the feasibility of our predictions. Significantly more than half of the predictions were accurately predicted (69%), whereas only 2 predictions (11%) were assessed to be false positives.

In the next stage of our study we will investigate the need for other types of cheap-to-acquire socio-technical information to refine our model. Also, further work will evaluate the method on a wider range of case studies. Finally, our method can potentially be used to identify opportunities for cost savings in existing systems, as well as predicting the costs and risks of new developments. The method may be also used to assess organisational readiness for various compliance programmes, such as clinical guidelines for management of chronic conditions like diabetes. The guidelines can be modelled as sets of data journeys that must be in place in order to comply with the guidelines, and organisational readiness can be measured in terms of how many of the needed journeys are in place at the start of the compliance programme.

Acknowledgements

This project is supported by funding from the UK EPSRC. We thank the Health eResearch Centre (HeRC) at the University of Manchester and the Farr Institute for providing access to the NHS case studies, and of course to the NHS clinical staff who wrote them.

Appendix A. NHS Case Study business processes

This section gives the business processes of a typical GP and hospital in the NHS when a GP patient might have a fracture and needs an X-ray scan to be taken at the local radiology department.

1. A GP fills in a request card to initiate the process of requesting a radiograph. The request card describes the type of X-ray needed and the patient's details. The request card is sent by post to the radiology department at the clerical reception area.

2. At the radiology department, a member of clerical staff receives the request card and creates an appointment for the patient in the radiology system. A letter containing the time and date of the appointment is created and sent to the patient through the post.

3. Before the patient arrives at radiology, the packet with the patient's previous X-rays and reports is transferred from the Filmstore area to the radiology clinical area by clerical staff using a trolley. If the patient has no previous X-ray scans, a member of clerical staff will create a packet at reception and takes it to the clinical area. A label with the patient's identification details will be attached to the packet.

4. On the day of the appointment the patient arrives at reception. Clerical staff will guide him to the clinical area. At the clinical area, a new X-ray is produced by a radiographer. The new X-ray is placed inside the patient's packet. The packet is then put into a pigeon hole by the clinical staff to be transferred to reception.

5. The packet is then transferred to the reception area by a member of clinical staff. The packet is then placed in a pigeon hole by clerical staff, from where a radiologist collects it. The radiologist takes the packet to his/her office, examines the X-ray scan and dictates a report onto a cassette.

6. The radiologist gives the cassette and the packet to the secretary who transcribes the report into the radiology computer system. The report is printed and given to the

radiologist to verify. If changes have to be made, the secretary amends the report in the system and prints it for verification.

7. A print out of the final report is placed in the packet by the secretary. The packet is then placed on a trolley to be sent back to the Filmstore by clerical staff. The secretary prints another copy of the report and puts it into an envelope to be sent to the GP. The porter collects the envelopes and transfers them to the porters area into a pigeon hole based on the GP address. The courier collects the envelopes from the pigeon hole and transfers them to the GP reception. The GP secretary gives the reports and the patient's folder to the GP. Sometimes, the GP secretary scans the printed report and inputs it into the GP system. The scanned report is linked with the patient's record. The GP accesses the scanned report.

Appendix B. Constructing the Data Journey Model

This section describes the process of designing the data journey model of the NHS case study. The model represents the journey of data needed by a GP to decide on an action plan when a patient may have a fracture, based on the business processes given in Appendix A. The steps shown here follow the bottom-up approach described in section 3.2.

Step 1: Identify data entities of interest

The first step after understanding the process and identifying the scope of the movement is to identify the data entities of interest. These are the data we want to move to the new development and their transformations. They can usually be derived from the scope of our journey. The data entities from the NHS case study are the data that a GP needs to decide on an action plan. These are the patient's identification details and the radiograph findings (referred to as report).

However, in order to create a report, a radiograph image (X-ray) is needed. But, what initiates the process of taking an X-ray? Data, once created, can be transformed, annotated, and updated before it is used by a consumer. In order to track the flow of moving data we

need to trace previous forms of that data to find its origins. For example, a GP has to request an X-ray to be taken by filling in a request card and sending it to the radiology department of the foundation trust. The request card will then cause an appointment to be made for the patient to attend the radiology, and the X-ray is taken.

Step 2: Identify the data containers in which data entities are stored.
Once we have identified the data of interest, we have to find the containers from which those data originate, are moved into and are finally made use of. Containers are stable, non-transferable places in which data can be stored. Containers can be electronic databases or physical locations, such as desks, filing cabinets or even pigeon holes. Data containers we identified from the NHS case study:

- GP's desk
- GP reception desk
- Radiology clerical reception desk
- Radiology information system's database
- Patient letter box
- Filmstore storing area
- Radiology clinical desk
- Radiology clerical reception pigeon hole
- Radiologist's desk
- Secretary's desk
- Porter area pigeon holes
- GP system database

Steps 3 and 4: Identify the routes and the media by which data are transferred.
After we find the containers of the journey, we identify the routes and the medium by which data are transferred from a source container to a target container. The medium is the means by which data is moved and can be in electronic or physical form, such as a sheet of paper, a request card, a folder, a label, etc. The routes, medium and the data entities moved we identified in the case study are:

- The request card is transferred from the GP's desk to the reception desk, and lastly to the radiology clerical reception desk. It is in physical form and contains the following data entities: the patient's NHS ID*, patient demographics, type of X-ray request.

- The radiology department patient packet is moved from the filmstore to the clinical desk, to the clerical reception desk, pigeon hole, radiologist's desk, secretary's desk and finally back to the filmstore area. It has a physical form and contains the data entities of: the unit ID, patient demographics, previous X-ray images, previous reports, new X-ray, new report.

- The cassette is transferred from the radiologist's desk to the secretary's desk and contains the data entities: patient identification details (various specified depending on the preferences and habits of the radiologists) such as, NHS or unit ID, name, surname, date of birth, the report. (A single cassette usually contains multiple dictations reporting on numerous patients.)

- The details captured in the cassette are moved into the radiology's system database which contains: the patient's demographic information, address, telephone, GP details, next of kin, etc.

- The report is moved from the radiology system database to the secretary's desk and it contains: the NHS and unit ID, patient name, surname, date of birth, radiograph findings.

- The envelope moves from the secretary's desk to the GP reception, and then to GP system database. It contains the patient's details and the report.
- The report is moved from the GP reception desk to the GP system database.
- The GP patient folder is moved from the GP filing cabinet to the reception desk, and finally to the GP's desk. It contains all the details of the patient since first registered with the GP and the report.
- **Note:* Each patient has a unique NHS ID. The NHS ID is given to the patients when they are born or become eligible for NHS care. When a patient attends hospital, they get a hospital ID, called unit ID. The Unit ID is unique per patient per hospital. Hospitals use the unit patient ID, but GPs usually use the NHS ID.

Step 5: Identify the actors interacting with containers.

The fourth step in constructing a data journey model is to identify the actors who interact with the previously identified containers to create, use or transform data entities stored in them. Actors can be people or systems and interact with the data stored in a container. They do not interact with data while it is moving between containers. The actors we identified in the case study are the following:

- GP
- GP secretary
- Patient
- Radiology secretary
- Radiology clerical staff
- Radiology clinical staff
- Radiologist

- Radiographer

Step 6: Draw the data journey diagram

The final step is to diagrammatically represent the data journey model using the notation given in figure 1 on page 8. The result of the representation is illustrated in figure 7 on page 28.

Appendix C. Other data journey boundary diagrams of the NHS case study

Once we have created the data journey model of the data entities of interest, we can overlay on it the boundaries; socio-technical information to help us identify places of the journey of high costs and risks.

Figure C.10 gives the organisational boundaries, and figure C.11 the media boundary. Figure C.12 shows the actors interacting with the containers to create, consume or transform data in order to produce some value. All figures note journey legs that crossed a boundary with a red warning sign indicating a likely place of high cost.

References

[1] M. Jørgensen, What we do and don't know about software development effort estimation, IEEE Software 31 (2) (2014) 37–40.

[2] J. S. Ash, D. W. Bates, Factors and forces affecting EHR system adoption: report of a 2004 ACMI discussion, Journal of the American Medical Informatics Association 12 (1) (2005) 8–12.

[3] R. Heeks, Health information systems: Failure, success and improvisation, International journal of medical informatics 75 (2) (2006) 125–137.

[4] M. Berg, Implementing information systems in health care organizations: myths and challenges, International journal of medical informatics 64 (2) (2001) 143–156.

[5] T. Greenhalgh, H. W. Potts, G. Wong, P. Bark, D. Swinglehurst, Tensions and paradoxes in electronic patient record research: A systematic literature review using the meta-narrative method, Milbank Quarterly 87 (4) (2009) 729–788.

[6] H. Wagner, O. Pankratz, W. Mellis, D. Basten, Effort of EAI Projects: A Repertory Grid Investigation of Influencing Factors, Project Management Journal 46 (5).

[7] A. L. Mark, Modernising healthcare—is the NPfIT for purpose?, Journal of Information Technology 22 (3) (2007) 248–256.

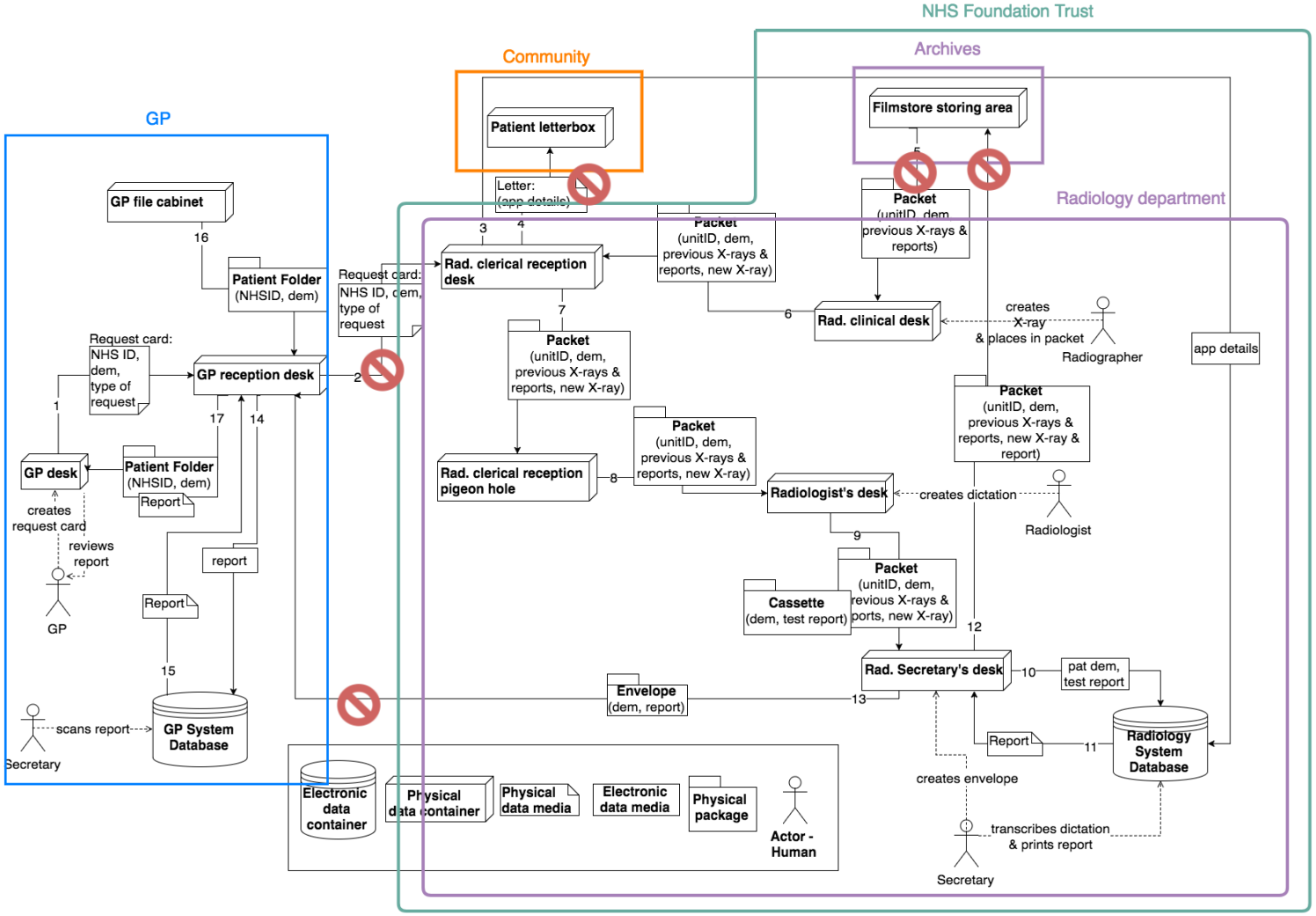


Figure C.10: Data journey model of the old system with overlaid organisational boundaries.

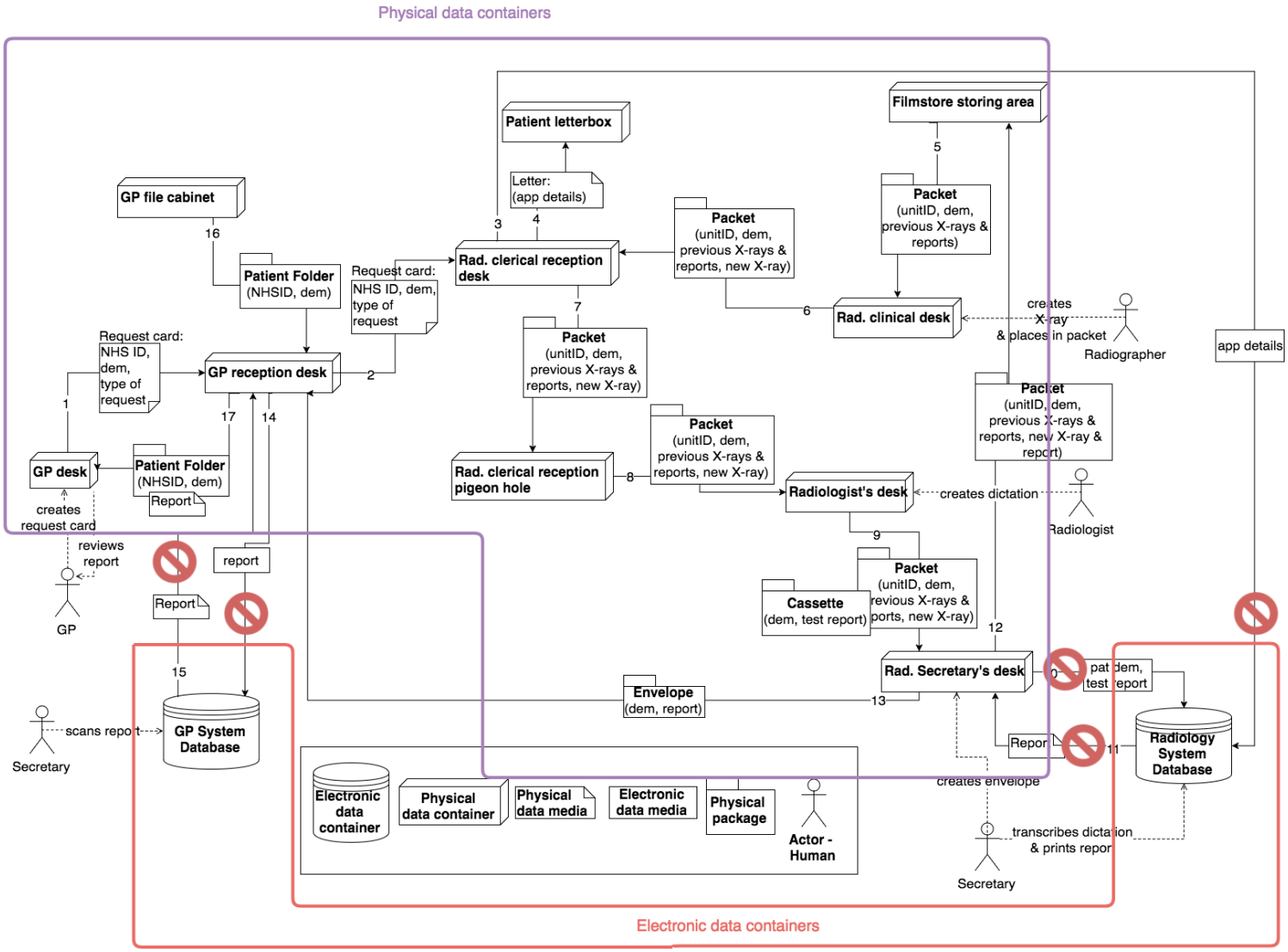


Figure C.11: Data journey model of the old system with change of media boundary.

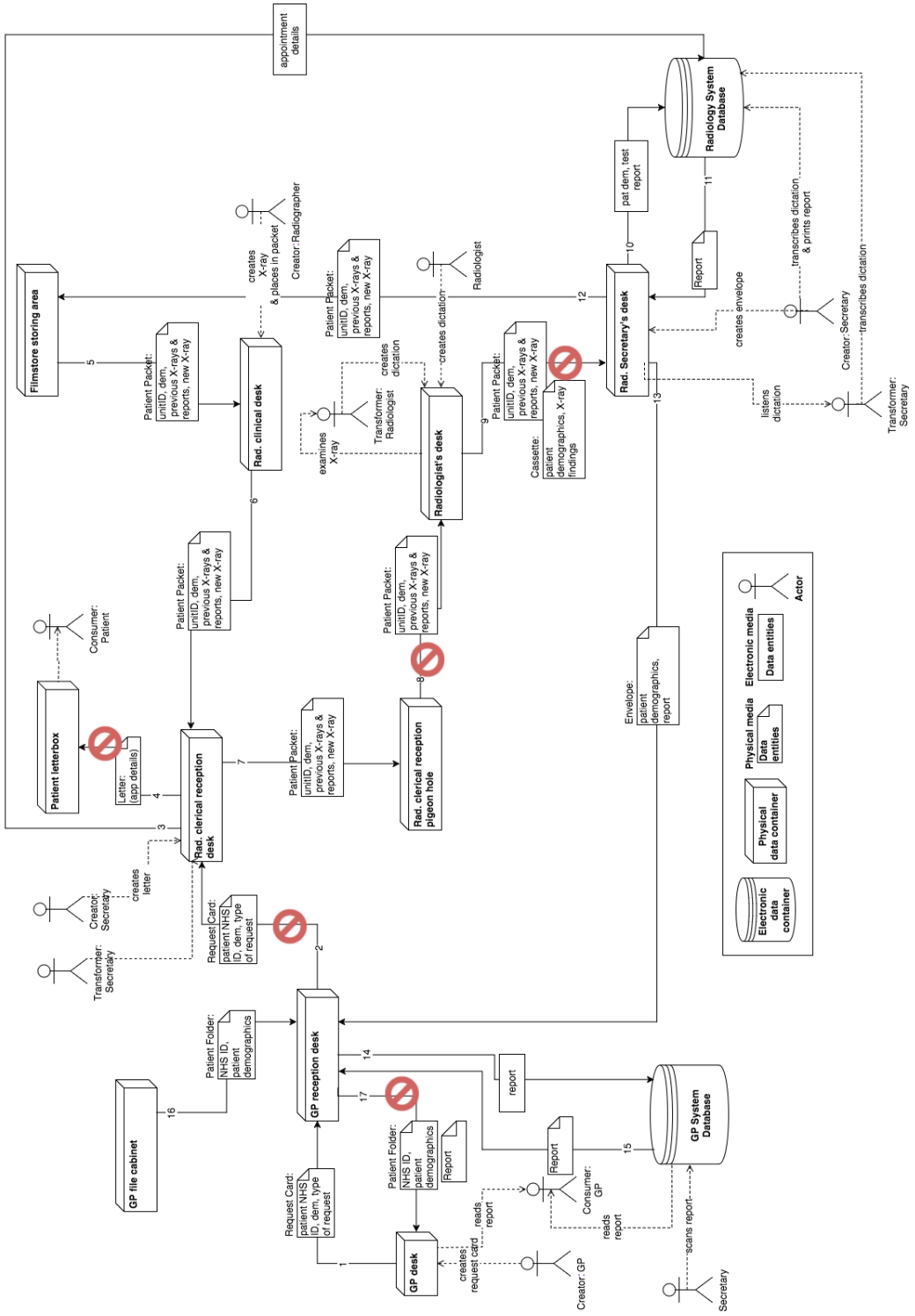


Figure C.12: Data journey model of the old system with actors' role boundary.

[8] J. Hendy, B. C. Reeves, N. Fulop, A. Hutchings, C. Masseria, Challenges to implementing the national programme for information technology (NPfIT): a qualitative study, Bmj 331 (7512) (2005) 331–336.

[9] W. L. Currie, M. W. Guah, Conflicting institutional logics: a national programme for it in the organisational field of healthcare, Journal of Information Technology 22 (3) (2007) 235–247.

[10] M. Y. Becker, Information governance in NHS's NPfIT: a case for policy specification, International Journal of Medical Informatics 76 (5) (2007) 432–437.

[11] N. A. Office, E-Borders and Successor Programmes, Tech. rep., Home Office, UK (2015).

[12] B. Boehm, C. Abts, S. Chulani, Software development cost estimation approaches - a survey, Annals of software engineering 10 (1-4) (2000) 177–205.

[13] M. Jørgensen, S. Grimstad, Software development effort estimation—demystifying and improving expert estimation, in: Simula Research Laboratory, Springer, 2010, pp. 381–403.

[14] E. Mendes, Effort and risk prediction for healthcare software projects delivered on the web, in: Practitioner's Knowledge Representation, Springer, 2014, pp. 107–122.

[15] A. Trendowicz, Software Cost Estimation, Benchmarking, and Risk Assessment: The Software Decision-Makers' Guide to Predictable Software Development, Springer Science & Business Media, 2013.

[16] A. Trendowicz, R. Jeffery, Principles of effort and cost estimation, in: Software project effort estimation, Springer, 2014, pp. 11–45.

[17] P. P.-S. Chen, The Entity-Relationship Model—Toward a Unified View of Data, ACM Transactions on Database Systems (TODS) 1 (1) (1976) 9–36.

[18] C. Batini, S. Ceri, S. Navathe, Entity Relationship Approach, Elsevier Science Publishers BV (North Holland), 1989.

[19] R. S. Aguilar-Saven, Business process modelling: Review and framework, International Journal of production economics 90 (2) (2004) 129–149.

[20] Y. L. Chen, et al., Data flow diagram, in: Modeling and Analysis of Enterprise and Information Systems, Springer, 2009, pp. 85–97.

[21] R. Becker, S. G. Eick, A. R. Wilks, et al., Visualizing network data, Visualization and Computer Graphics, IEEE Transactions on 1 (1) (1995) 16–28.

[22] M.-M. Bouamrane, A. Rector, M. Hurrell, Using ontologies for an intelligent patient modelling, adaptation and management system, in: On the Move to Meaningful Internet Systems: OTM 2008, Springer, 2008, pp. 1458–1470.

[23] E. S. Yu, Social modeling and i*, in: Conceptual Modeling: Foundations and Applications, Springer, 2009, pp. 99–121.

[24] M. M. Yusof, J. Kuljis, A. Papazafeiropoulou, L. K. Stergioulas, An evaluation framework for health information systems: human, organization and technology-fit factors (hot-fit), International journal of

medical informatics 77 (6) (2008) 386–398.

[25] J. Rumbaugh, I. Jacobson, G. Booch, Unified Modeling Language Reference Manual, The, Pearson Higher Education, 2004.

[26] Y. L. Simmhan, B. Plale, D. Gannon, A survey of data provenance in e-science, ACM Sigmod Record 34 (3) (2005) 31–36.

[27] T. Pardo, A. M. Cresswell, S. S. Dawes, G. B. Burke, et al., Modeling the social & technical processes of interorganizational information integration, in: Proceedings of the 37th Annual Hawaii International Conference on System Sciences, IEEE, 2004, pp. 8–pp.

[28] S. Flowers, Software Failure: Management Failure: Amazing Stories and Cautionary Tales, John Wiley & Sons, Inc., New York, NY, USA, 1996.

[29] R. N. Charette, Why software fails, IEEE Spectrum 42 (9) (2005) 42–49.

[30] I. Eleftheriou, S. Embury, A. Brass, Data journey modelling: Predicting risk for IT developments, in: Proceedings of the 9th IFIP WG 8.1. Working Conference on the Practice of Enterprise Modeling, PoEM 2016, Springer, 2016, pp. 72–86.

[31] S. W. Ambler, Process patterns: building large-scale systems using object technology, Cambridge University Press, 1998.

[32] D. Budgen, Software design, Pearson Education, 2003.

[33] A. Koenig, Patterns and antipatterns, Journal of Object-Oriented Programming 8 (1) (1995) 46–48.

[34] J. L. Vann, Resistance to change and the language of public organizations: A look at "clashing grammars" in large-scale information technology projects, Public Organization Review 4 (1) (2004) 47–73.

[35] I. Eleftheriou, S. Embury, A. Brass, Light Touch Identification of Cost/Risk in Complex Socio-Technical Systems, in: Proceedings of the 10th IFIP WG 8.1. Working Conference on the Practice of Enterprise Modeling, PoEM 2017, Springer, 2016, to appear.