Technical Report: 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 a new system development is 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 new software systems that will reuse or extend their existing information infrastructure.

We analysed case studies describing new system developments undertaken by providers of health care services, looking for common points of risk and high cost. The results highlight 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 a new methodology aiming to predict places of high cost and risk when existing data needs to move to a new development. The methodology is lightweight and combines technical and social aspects, but relies only on information that is likely to be already known or will be cheap to acquire. We use/propose the data journey model, an approach of modelling data movement across organisations, as the central component of our method.

To assess the effectiveness of our methodology, 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 drive organisations to develop new, more advanced information systems (IS) to share and integrate their information. But, realising value from these new IS require hard decisions to be taken; is the new system development or re-design worth making? Is the value to be gained more than the costs of developing and maintaining the new development?

Predicting the success or failure of new developments is hard. There is a complex mix of a variety 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 harder to predict challenges stem from the social aspects of the organisation; its people, policies, processes, governance, etc.

Costs arising from technical and social barriers are often underestimated. An example is the National Programme for IT (NPfIT), an initiative by the Department of Health in the United Kingdom aiming to use modern information technologies to improve the delivery of health services and the quality of patient care [1]. Numerous information sharing and integration solutions were introduced under NPfIT, but after 12 years and a total forecasted cost of 9.8 billion UKP, the planned central, integrated system has yet to be established [2, 3, 4]. Another example is the e-borders programme of the Home Office in the UK. The programme was initiated on 2003 and aimed to collect information 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 [5].

In addition to the ruggedness of predicting failure or success of IT systems, existing methods don't ease the decision making process. Despite a plethora of modelling techniques and notations found in the literature for use during information systems design [6, 7, 8, 9, 10, 11, 12, 13], we only found a handful of methodologies that give equal prominence to both social and technical factors [12, 13, 14]. Of these, none were sufficiently lightweight to be used in early stage go / no go decision making.

To aid the decision making process, we devised a lightweight technique, called data journey modelling, that captures the movement of data through complex networks of people and systems [15]. To create our model, we analysed a collection of 18 case studies written by staff working for the NHS (National Health Service) in the UK, describing factors that contributed to the failure or success of recent IT developments, in which the case study authors had been involved in. The results 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. [16, 17]).

*I am 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)

In this paper, we propose a methodology that uses the data journey model to identify high risk data movement places of planned developments that may introduce unforeseen costs. Our contributions are:

- A lightweight method that combines social and technical information to predict barriers to data movement that can impose high costs, based on the data journey model (Section 3).
- The application of the data journey model to a real world case study from the NHS domain, describing data moved from a General Practitioner (GP) organisation to the radiology department of a foundation trust (Section 4).
- A retrospective evaluation of the methodology to assess the precision of the predictions with the help of NHS domain experts (Section 4.3).

The results of the evaluation are promising, showing that our methodology can identify barriers to data movement and can potentially predict costs and determine corrective actions to avoid unnecessary costs and risks. The following sections describe the methods we used to develop our methodology (section 2), present the methodology (section 3), and next we provide the details of the retrospective evaluation conducted (section 4). Finally, section 5 discusses the results of the evaluation and conclusions drawn.

2. Methods

To develop our methodology it was necessary to understand what elements of an IS infrastructure lead to high cost and risk, and to devise a method that could look for early warning signs of the presence of such elements, based on information that can be cheaply and quickly acquired. We took the following approach to the development of our method:

1. We examined a set of 18 case studies from the NHS domain, to understand the factors that were present when IT failure of some kind occurred. From each case study, we extracted the social and technical elements present that, according to the authors of the case studies, contributed to the failure.
2. From the results of step 1, we formulated a set of data movement patterns that under some criteria introduce high costs to the development of the IT projects.
3. Based on the patterns, we devised a modelling approach that captures just the elements identified as being significant factors.
4. We then developed a methodology that uses the model and overlays social and technical information captured to predict costs and risks.
5. Finally, we performed a retrospective evaluation of the method that resulted, using it to predict areas of high cost and risk in a Radiography department of a UK hospital. Our predictions were compared with the actions taken by hospital staff to reduce the cost of one of the primary processes undertaken by the department.

The case studies used for our research were written by NHS staff taking the "Informatics for Healthcare Systems" professional development course at the University of Manchester, during the 2013 academic year. They describe a variety of IT developments in the NHS domain 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, while describing 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.

We examined the case studies to identify factors contributing to the failure of the IS, both technical and social. The most common factors contributing to IT failure as identified by the authors of the case studies are related to people: for example, staff resistance to changing their processes, insufficient stakeholder engagement in decision making, lack of shared vision, etc. Other factors of technical and organisational nature were also identified. Examples include conflicting data formats, disconnected data silos, heterogeneous sources, inadequate system performance, governance issues, ethics, politics and others. All 32 factors identified are further described in another paper [15].

From the contributing factors identified above, we found that movement of data either between systems, organisations, or even people, is a key indicator of cost and possible failure. Examining data movement examples from the case studies we retrieved patterns that, if present in a new IT development under some specified conditions, can be an indicator of high costs [15]. From the patterns we found that discrepancies between some key properties of the source and target of the movement can introduce costs to the new development. For example, if a source system stores data in a physical form, but a target system requests it in electronic form, then a transformation cost will be imposed in either side of the movement. Similarly, if a source belongs to a different organisation than the target system, then some governance or ethical issues may appear.

However, data sharing is of vital importance in providing the right information to the right people at the right time, and we can't eliminate all data movement instances just because is risky. Hence, we developed a method that, based on the patterns, predicts early warning signs of costly movements of data of a new IT development. The following section begins by presenting the data journey model which captures the journeys taken by a set of data through an information infrastructure, and then proposes a methodology to help us predict parts of the journey that can potentially introduce high costs to the new development.

3. Identifying barriers to data movement

Analysing the case studies we found that data movement, whether between systems or people, is a potential indicator of the presence of costs and risks in a socio-technical system [15]. To predict points of cost / risk in advance, we need a way of identifying the places in an information infrastructure where data is moved between two entities which differ in some way significant to the interpretation of the data. These are the places where the "portability" of the data (the ability of information to survive 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.

Hence, we need a methodology that assists in:

- Modelling the necessary parts of the existing information infrastructure including the places where data is stored, and the links between these places that enable the sharing of data.
- Modelling the movement of data from a point of entry in the infrastructure to the point of use by a new consumer (human user or a new IT development).
- Adding cheap to acquire information (social and technical) on the model to identify key heterogeneities between the elements of the infrastructure that can potentially impose high costs and risks to the movement.

3.1. Our proposed methodology

Our methodology begins by identifying the set of data that are needed to the new development. We then model the journey of the data from its original location within the information infrastructure of the organisation to the new development, using the data journey model. Once we have modelled the data journey we add to the model the boundaries; the information we retrieved from the patterns to help us identify discrepancies between key elements of the organisation. Finally, we predict points of high cost by identifying the places of the journey in which data have to move from a source to a target that belongs to a different type of boundary.

In the remainder of this section, we present the model and notation we have designed to capture the first three requirements: the infrastructure and movement of data. In section 3.4, we describe how this model is overlaid with additional information that can help us to discriminate between movement steps that are unlikely to cause problems, and those where significant heterogeneities exist between the producer and the consumer of the data. In section 3.3, we propose a process for creating and using the resulting models, to predict points of cost and risk in a new development.

3.2. Data journey model

To guide the design of the model, we use the metaphor of a "journey" to represent the movement of a set of data entities between organisations. Journeys are purposeful, implying that the data is needed at its destination for some value-creating step. For example, the movement of a blood analysis request card from a GP organisation to the pathology lab, is a data journey that initiates the production of test results.

The aim of the data journey model is to provide a lightweight way to conceptually represent the movement of data within or across organisations. It doesn't attempt to provide a complete representation of the organisation, or its processes, nor does it model the complete set of data movements, but just the journeys of the dataset of our interest. The model is used in an agile way, showing a simple set of possible information that will help us identify

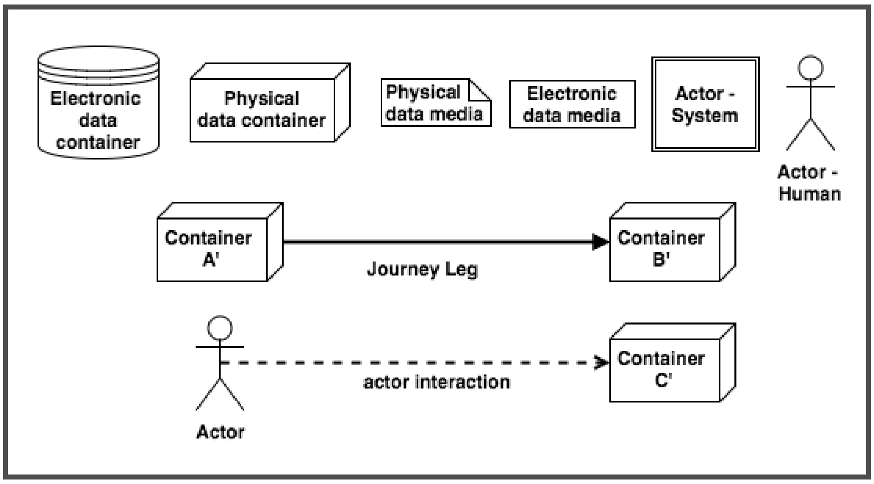

Figure 1: The notation used when modelling a data journey.

more than just the technical issues related with the data flows of an organisation. The emphasis of having a lightweight model 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. Thus, we designed a model that captures:

- The journey of data across an information infrastructure, including the entities in which data is stored (physical and electronic) and the routes by which data moves between them.
- The actors using, creating or transforming the data moving in the infrastructure, both human and technological, that can affect the journey.
- The points at which key heterogeneities in the interpretation of data occur that can potentially impose costs to the journey.

3.2.1. The Landscape

The *landscape* of the model captures the information infrastructure in which data flows. It may represent a single organisation or a more complex network of systems of collaborating organisations. An organisation is viewed as a collection of entities (people and systems) between which data moves along well-defined routes. The landscape captures the places in which data rests when it is not moving, and the routes between them along which the data can move.

For example, when a GP requests blood test results from the pathology lab of a hospital, data needs to travel from the GP secretary's desk (in the form of a request card and blood sample), to the hospital porter's pigeon holes, 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.

Data containers are the resting places that store the data of our interest. 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 icon and physical ones with a rectangular box as shown in figure 1. In the lab pathology example described above, containers are the GP's desk, and the GP reception desk (storing the request card and blood sample), the pigeon holes of the hospital, the pathology lab secretary's desk, and the pathology system database.

Data stored in a container can travel to another container through some already established route. In the data journey model we call them *links*. A link connects two containers if there is an existing connection 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. We denote links with a straight line connecting the two containers. The direction of the link shows the movement of data from the source to the target container, as shown in figure 1.

Each link transports data using a specific *medium* (paper, electronic, X-ray films, cassette, etc). Media can be of physical or electronic form. For example, the blood test request is moved from the secretary's office to the pathology lab on a piece of card posted in an envelope. The test results move from the lab's database to the GP's system through an internet connection.

Actors are the people or IT systems that interact with the containers to create, consume, or transform the data resting at it. Actors are denoted using the actor symbol of the UML notation [18], and the interaction with the containers is shown with a dotted arrow beginning from the actor and ending with the container with which interacts. Several actors can interact with one container, but each interaction is explicitly modelled. One actor can also have several interactions of different actions with the same container.

3.2.2. The Data Journey

Having modelled the existing infrastructure through which data is currently moved, we can now model new movements of data needed to meet any planned new requirements. We call the new movement the *journey* that we want to evaluate risks and costs for.

Journeys involve movement from a point of origin (the point at which the data first enters into the scope of the model) to a final destination (the point at which the data is used for its purpose), passing along the way through zero or more way points, at which the data may "rest" for either a short or long time before continuing its journey to the end. Those way points may themselves be producers or consumers of this and other data items, or they may merely hold data. Journeys may also be short (within a department or organisation) or long (covering a network of cooperating organisations). They may describe movement that has happened in the past (in an existing infrastructure) or they may describe planned movement for the future (a proposed new IS development).

Sometimes a direct sequence of links between the source and target container doesn't exist driving the data to move through intermediary containers using existing links. Those links are called *journey legs*. A set of consecutive journey legs create the route of the journey.

The data journey meta model is given in another paper [15].

3.3. Modelling a Data Journey

To apply our methodology we first have to identify the scope of the movement we want to model, the set of data to be transferred to the new development and the business processes in place that move the data within or between organisations. For example, a new IT system is needed that will use data already existing in another source, or a new guideline requires the sharing of data to an external organisation. Having defined our scope, we identify the elements of the infrastructure needed to accomplish the movement within our scope (data entities, containers, actors). We then must have a clear idea of the business processes needed to move the data by consulting a domain expert.

Based on the bottom up approach, we start the modelling process by identifying the elements of the model which are within our scope by:

1. Firstly, we identify the data entities of our interest.
2. Secondly, we list the containers in which the data entities of our interest originate from, are moved into and finally the target in which will be placed.
3. Thirdly, we identify the medium in which data are transferred from a source to a target container.
4. Then, we list the actors interacting with the containers to create, transform or consume data entities stored.

Once we have identified the elements of the journey of our interest we can start designing the model using the notation given in figure 1. Below steps are a guideline following the bottom up approach. An example illustrating the steps below is given in figure 2.

1. We start by creating the containers. We make a rectangular box for each physical container and a cylinder for each electronic container and we name them accordingly.
2. We then connect the containers to form the journey legs using a straight line arrow starting from a source container and ending to a target container based on the direction of the movement of the data described in the business processes. We then number the arrows for future reference.
3. For each journey leg we identify the data entities being moved and design a rectangle for data moved using an electronic type of media and a rectangle with a folded angle for data moved through a physical type of media. In the middle of each media symbol we write the data entity being moved by the media.
4. We then add the actors which create some value in the scope of our journey i.e. creating a data entity, transforming it, or using it. We create a UML actor symbol for each human actor and a doubled line rectangle for a technical actor. We name each actor with its business position (e.g. GP, clerical staff, clinical staff, informatician, EPR system) and role in the journey (creator, transformer or consumer of data).

5. We then connect each actor with the container it interacts with a dotted arrow, starting from the actor and ending to the respecting container. We name the arrow with a small description of the interaction (e.g. inputs data, creates report, analyses blood sample, etc).

Having followed above steps we now have the data journey diagram of the movement of our interest. The last part of the figure 2, shows the journey of the lab pathology example given earlier in this section. The following section describes the process of overlaying information on top of the data journey model to identify parts of the journey with high costs.

3.4. Identifying barriers

Once we have created a data journey diagram of the movement, we can identify parts of the movement that, because of some type of barrier, can introduce some costs to the movement or new development. From our analysis of the 18 case studies described in section 2, we found that a key indicator of high costs of data movement is the substantial heterogeneities between the elements of the data journey.

We hypothesise that a journey leg has a barrier if it has high cost imposed to it. A barrier is a discontinuity, a difference in the properties of the source and target containers of a journey leg, that causes a high cost to the movement.

Discontinuities can be caused by numerous boundaries; factors that can affect the movement of data and impose barriers to journey legs. An example is the organisational structures in which data are moved. Moving data from an organisation to another we introduce the cost of governance and regulation compliance. Putting the organisational structures on top of the data journey diagram we can identify the journey legs that cross this boundary. and have the barrier of complying with the regulations of the organisation before moving the data. Organisational structures is an easy and cost effective information to acquire. Hence, by applying them on the data journey we can identify barriers with cheap information without using valuable time of the clinicians 3.

Another factor could be the nature of a container; either electronic or physical. Moving data from a physical container to an electronic one causes the cost of manual input of data into the system by either transcribing or scanning. On the other hand, moving data from an electronic to a physical container causes the costs of printing.

Other dimensions can be the level of technological familiarity of the staff, types and formats of the data entities, different application program interfaces (APIs) of the IS used, geographical locations of the containers and other social and technical dimensional factors that can impose costs to the movement. A warning sign exists when the source container of a journey leg belongs to a different boundary than the target container of the leg. We hypothesise that each journey leg with a warning sign can impose costs to the journey of data.

4. Evaluation

Given the nature of our methodology is very difficult to evaluate it in a real world example. Ideally, we would need an organisation in need of a system development or re-design and 2 teams of software engineers; one to apply the cost saving changes predicted by our model and another one to change the current system with respect of changes identified by domain experts. Then, we would wait until the two teams finish implementing the systems to compare cost savings of both systems and evaluate whether our methodology accurately predicted barriers. However, this type of evaluation needs resources and time that we don't have. Instead, we are conducting a retrospective evaluation of a real world case study. We identified a case study in which a new system development was recently introduced to the organisation. We applied our methodology to the old system before the introduction of the new information system and predicted barriers of high costs and risks places. We then model the new system to find changes made and compare the two models to assess the accuracy of our predictions.

4.1. Success criteria

We defined a set of success criteria to assess the success of our model in the chosen case study. We don't expect our model to predict all the changes made by the domain experts in the new system since that would require detailed and heavy research, plans and effort. Whereas, we are evaluating whether our lightweight model can cheaply and quickly predict some of the high costs places where savings can be made without investing extreme resources (time, effort, money). Our model will be successful if it is lightweight and accurate-enough.

We define the lightweight property of the model in terms of the time and effort taken to produce a model and have a set of predictions. We measure time in terms of the time invested by staff of the organisation to provide us domain information needed to make the model. Effort is measured in terms of the time required to create the actual model and predict places of high costs. We set the threshold of staff's time and time taken to create the model to 1 working day (5 to 7 hours) each.

In order to assess the accuracy of the model we assume that the new system introduced by the domain experts has a correct set of changes. However, we also assume that there are other places of potential cost savings that were not changed by domain experts for various reasons, i.e. limited time and resources. Hence, we consider a prediction to be accurate if is solved in the new system or another domain expert assess it as a sensible and feasible change that would save costs.

Since we have a simple and lightweight model we don't expect all predictions to be accurate. However, every inaccurate prediction has the accumulative cost of further investigating it. We are considering the model to be accurate if the wrong predictions are less than the correct predictions. A prediction will be assessed as correct if has been solved in the new system or wasn't solved, but a domain expert assessed it as sensible and feasible.

4.2. Application on NHS case study

To evaluate our methodology we used a case study from a UK NHS foundation trust, other than those analysed to develop our methodology. In this case study we examined a data movement example from the radiology department of the Trust. The radiology department provides a wide range of diagnostic imaging services. Specifically, the case study describes the movement of data that a General Practitioner (GP) needs to decide on an action plan when a patient may have a fracture. When a GP considers that a patient may have a fracture, he/she requests an X-ray to be taken at the local hospital's radiology department. When the patient arrives at the radiology department, a radiographer takes an X-ray image of the patient. A radiologist reviews the image and dictates a report with his findings. Then, the secretary transcribes the report into the system, prints and sends it to the GP to decide on an action plan.

As with any large commercial or governmental institutions, some aspects of the details of the case study are confidential, and we do not have permission to publish them. In this report, we illustrate our ideas with a scenario inspired by the nearby foundation trust case study. Although the models used to evaluate our methodology are retrieved from the actual case study, in this report we present and use a more general model, typical of those used in a range of NHS hospitals. The results of the evaluation presented are the original ones produced from the actual case study.

4.2.1. Methods

To create the journey model of the radiology data, we conducted semi-structured interviews with clinicians working at the radiology department of the foundation trust. From the interviews we gathered domain knowledge needed to create the model, like organisational structures, available information systems and the business processes of taking an X-ray, and creating a report. We developed the model in an agile-iterative process gathering continuous feedback from the clinicians.

Following subsections describe the process of modelling the journey of the data in the old system, the new system in place and the results of the comparison.

4.2.2. Modelling the data journey of the old system

We begin applying our methodology by identifying the set of data that we want to transfer to the new development, and the scope of the movement. The scope of the movement in the NHS case study is:

> *To model the journey of data needed by a GP to decide on an action plan when a patient may have a fracture.*

To start modelling the journey we must have a clear idea of the processes that move data between the two organisations; GP and hospital. Appendix A describes the business processes in which data moves from the GP to the radiology department of a hospital based on the old way. A step by step application of our methodology is given in Appendix B. The data journey diagram of the old system is illustrated in figure 4. Although our evaluation is based on data captured by a UK F.T., both appendices and the diagram describe a general movement of data, typical of those used in the radiology department of NHS trusts.

4.2.3. Identifying barriers

To identify the places that can introduce some costs to the movement we add to the data journey diagram the factors we think can affect the movement of the data. These are the organisational structures, actors and data media described in section 3.4. We derived information on the factors from the semi-structured interviews with the clinicians. We then overlay them on the data journey diagram. Figure 5 shows the organisational boundary layer on the data journey of the old system. The actors and data media boundary figures are given in Appendix C.

The next step of our methodology is to identify the journey legs which cross the boundaries. Whenever a journey leg crosses a boundary a barrier may exist which can introduce costs to the leg. In figure 5 legs that cross a boundary are illustrated with a warning sign.

By combining all three boundary layers on the data journey model we have a heat map highlighting the journey legs with higher risk of imposing costs to the journey. Figure 6 shows the heat map of the three layers. From the figure we see the journey legs with number 4, 11, 13, 14, and 15 to have double line arrows representing two warning occurrences. The single line red bold arrows show the occurrence of one warning sign from either of the three boundary layers. Table 1 further describes the predicted barriers and gives potential causes.

Journey Leg No	Predicted barriers	Barrier cause
2	organisational and actors	Data moved between two organisations, GP and Foundation Trust. Also, data created were used by another actor of different position role.
3	data containers and actors	Data moved from the physical container of clinical reception desk to the electronic container of the radiology's system database. Also, data created by the secretary are used by an actor of different role; the radiology system.
4	organisational and actors	Data moved between two organisations, the FT and the community. Also, data created by the secretary of the FT are used by a user of different role; the patient.
5	organisational	Data are moved from the archives department of the FT to the radiology department.
8	actors	Data created by the Radiographer are used by another actor, the Radiologist.
9	actors	Data created by the Radiologist are used by another actor, the Secretary.
10	data containers and actors	Data moved from the physical container of radiology secretary's desk to the electronic container of the radiology's system database. Also, data created by the secretary are used by an actor of different role; the radiology system.
11	data containers	Data moved from the electronic container of the radiology's system database to the physical container of the radiology secretary's desk.
12	organisational	Data moved from the radiology department of the FT to the archives department.
13	organisational	Data moved from the radiology department of the FT to the porters area.
14	organisational	Data moved from the FT organisation to the GP organisation.
15	data containers and actors	Data moved from the physical container of the GP reception desk to the electronic container of the GP's system database. Also, data created by FT radiology secretary are used by the GP's system.

16	data containers	Data moved by the electronic container of the GP's system to the physical container of the secretary's desk.
18	actors	Data created by the GP's system are used by an actor of different role, the GP.

Table 1: Barriers predicted by our methodology.

4.3. Evaluation results

We applied our methodology to the old system FT and retrieved a set of predictions of potential high cost places. To evaluate our methodology and the accuracy of our predictions, we compare them with changes made to the new system by domain experts. Specifically, we will:

1. compare the old data journey with the new journey to find a set of changes that experts from the F.T. did.

2. Then, we compare the changes with our predictions to find any commonalities. If a high cost place was predicted by our methodology and was selected to be changed in the new model, we assess the prediction as accurate.

3. However, we ... (don;t know the full ground truth text) hence, we need the knowledge of domain experts to assess the feasibility of our predictions that haven't changed in the new system. Table 3: Feasible and sensible predictions.

4.3.1. Old and new system changes

The new system replaced the old X-ray machinery with a state of the art electronic equipment that stores X-ray images in digital form. X-ray images are uploaded to the Picture Archive Communication System (PACS) which replaced the old packages and the Archives department. The PACS system is integrated with the Computerised Radiology Information System (CRIS) responsible for receiving referrals, booking appointments, and managing patients. CRIS is fully integrated to key hospital information systems such as Patient Administration System (PAS), Order Communications and Electronic Patient Records (EPR). However, although most of the radiology departments in UK provide digital services, some GP organisations still have no means to integrate with. In those cases referrals and requests are received through post on paper format.

To evaluate our methodology we assume that the radiology department provides both electronic and physical referrals from GP organisations. Figure 7 shows the data journey model of the new system in place.

The data journey models used are centred from the radiology department and the flows within the GP organisation may vary in other organisations. However, we used typical examples used widely in GP organisations across the UK.

4.3.2. Results

By applying the methodology on the NHS case study we identified 5 journey legs crossing an organisational boundary, of which only one didn't have any governance barriers. Based on the clinicians expertise, the leg with no governance barrier was the one to be eliminated since governance is one of the hardest and most complex barrier to eliminate. See table 3 for results.

To do: Discuss results and predictions:

x out of x journey legs that our methodology predicted to have a barrier assigned were identified and changed in the new system. x out of x journey legs that our methodology predicted have not changed in the new system. However, blah blah of ground truth. So ... but have been assessed as feasible by the domain experts.

x out of x journey legs predicted, haven't changed in the new system and domain experts assessed them as not feasible.

The new system has some changes that haven't been predicted by our methodology.

The methodology found some journey legs with no barriers that haven't changed in the new system.

Overall, the results are promising showing that our methodology can identify boundaries and barriers of specific legs of a data movement development and potentially determine a corrective action or an alternative the boundaries with the highest value and cost of the data journey.

Journey Leg No of the old data journey	Predicted as costly?	Changed in new data journey?	Domain expert view	Result
1	No	No		
2	Yes	No		
3	Yes	Yes		
4	Yes	No		
5	Yes	Yes		
6	No	Yes		
7	No	Yes		
8	Yes	Yes		
9	Yes	Yes		
10	Yes	Yes		
11	Yes	Yes		
12	Yes	Yes		
13	Yes	Yes		
14	Yes	Yes		
15	Yes	Yes		
16	Yes	Yes		
17	No			
18	Yes	Yes		

Table 3: Data journey old model predictions and comparison with the new system. (TP=True Positives, FP=False Positives, FN=False Negatives)

5. Discussion

In this paper, we analysed a set of 19 case studies from the NHS domain to identify barriers that can introduce high costs to the development of new systems. The results of the analysis indicate that although movement of data is imperative, it can be affected by several socio technical barriers that can impose high costs to the development of new applications. Given that these costs are often underestimated, we need a way to better identify and predict barriers of enterprise scale data movement.

This paper proposes a new model that conceptualises the journey of a piece of data through a network of systems and organisations, called the data journey model. We also developed a methodology that applies some cheap to acquire information on the data journey model to predict socio-technical barriers to new data movement technological solutions. We conducted a retrospective evaluation of our methodology on a real world case study from the NHS domain. The results of the evaluation show that our methodology can successfully predict some of the barriers found in this case study. Significantly more than half of the predictions were accurately predicted.

A criterion of the success of our methodology was the time and effort needed to collect information needed and predict some barriers. Although the more information invested in the methodology will provide more accurate results, time acquiring this information will be taken away of clinicians' valuable time. Results obtained indicate that our methodology is lightweight; it can achieve predictions even by inputting some cheap to acquire information like the organisational structures and the salary positions of staff members of an organisation.

Results of the evaluation indicate that our methodology can accurately predict some sociotechnical barriers to new data movement solutions in a lightweight, simple way. It acquires cheap information and quickly produces predictions of potentially high cost places.

> *"Costs are indicative of both 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 take 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 mis-communication, mis-understandings, 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)." - Domain Expert*

Our paper presents an innovative new way to conceptualise and visualise the journey of data among organisations. Other techniques and tools found in the literature representing data movement only describe the technical factors affecting the cost of data movement. Others, identifying and combining both technical and social information are not abstract enough. They require detailed information which is hard and time consuming to acquire. To the best of our knowledge, this is the first study that conceptualises the journey of data among organisations in an abstract way that help us apply some cheaply acquired information about the organisation to identify barriers.

Because of short time-scale we only tested our methodology on one case study. However, we evaluated it on a real world case study of one of the most difficult and complex organisations, the NHS. Although, we can't claim the validity of generalising our methodology, the case study chosen to evaluate it has several wide range properties. It consists of movement of data between two different organisations with proven hard and costly communication. Movement of data contains both physical and electronic links. These properties along with the promising results of the evaluation suggest the possible capability of our methodology to be applied in other case studies of different environments. Further work will evaluate the methodology on a wider range of case studies.

Also, the model used to identify the barriers of the movement of data is a very simple one containing only three types of information, organisational structures, the type of the containers used to store data (physical or electronic) and the salary position of the staff of the organisations. Although, the model used is a very simple one, uses information of the organisation that can be quickly and cheaply acquired. Results obtained by evaluating the methodology show that even such a simple model can result to some useful and accurate predictions. In the next stage of our study we will investigate other types of cheap information to expand our model.

6. Conclusion

In this paper we propose a new methodology to identify barriers of moving a piece of data among a network of systems and organisations that can impose high costs when developing new or re-engineering a technological solution. Evaluating our methodology on a real world case study we obtained promising results indicating the capability of our methodology to quickly identify socio technical barriers to data movement using cheap to acquire information.

Future work will involve the expansion of the methodology to also predict quantifiable costs and risks of the identified barriers. Predicted costs and risks will be assessed and compared with value to be gained by developing a new application to make a go / no go decision. This work is the necessary first step in a longer term project to provide a mechanism predicting cost and risks of data movement in organisations such as the NHS, with the aim of assisting in the making of go/no go decisions regarding new information system developments

Finally, our methodology can potentially be used to identify opportunities for cost saving in an existing system, as well as predicting the costs and risks of new developments. So that, it will assist managers and stakeholders of large organisations make better informed decisions on whether a new development is worth developing or not. Also, the methodology may be 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 to check whether the organisation follows or not. If the organisation does not implement a data journey guideline will show the cost of compliance to the organisation.

Acknowledgements

This project is supported by funding from the UK EPSRC. We thank the Health eResearch Centre (HeRC) of the University of Manchester and the Farr Institute for providing us 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 provides the 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 ***define*** to the radiology clinical area***define*** by clerical staff using a trolley. If the patient has no previous X-ray scans, a clerical staff will create a packet at the 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 packet. The packet is then put into a pigeon hole by the clinical staff to be transferred to the reception.

5. The packet is then transferred to the reception area by the 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 where examines the x-ray scan and dictates the report in a cassette.

6. The radiologist gives the cassette and the packet to the secretary who transcribes the report into the radiology 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 to be verified by the radiologist.

7. A print out of the final report is placed in the packet by the secretary. The packet is then placed at 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 ***define*** 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. OR: 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 a data journey model

This section describes the process of designing a data journey model of the scenario given in section 4.2. Below steps follow the bottom up approach and are all based on the processes described in Appendix Appendix A.

To start modelling the journey we must have a clear idea of the processes that move data between the two organisations; GP and hospital. Appendix A provides a detailed version of the business processes in which data moved from the GP to the radiology department of a hospital.

To model the journey of data needed by a GP to decide on an action plan when a patient may have a fracture.

Appendix B.1. 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 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 an X-ray to be taken? Data once is created, can be transformed, annotated, updated and edited before used by a consumer. In order to track the flow of moving data of our interest we need to trace the previous forms of those 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 send it to the radiology department of the nearby foundation trust. The request card will then cause an appointment to be made for the patient to attend the radiology. The X-ray ...

Appendix B.2. 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 finally the target in which are placed. Containers are stable, non-transferable places in which data can be stored and rest. Containers can be electronic databases or physical places like a desk, file cabinet or even pigeon holes. Containers can be derived from the business processes by asking the question "Where is data entity 'x' stored?". In the case of a physical data entity we are looking for a stable, not moving place which can hold a data entity.

Data containers from the NHS case study in order of use from the process described in Appendix A:

- GP's desk

- GP reception desk
- Radiology clerical reception desk
- Radiology system 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

Appendix B.3. Step 3: Identify the media through which data are transferred.

After we find the containers of the journey, we identify the medium in which data are transferred from a source container to a target. Medium is the mean that moves data and can be in electronic or physical form like a sheet of paper, a request card, a folder, a label, etc.

Types of medium identified in the case study:

- Packet medium contains: unit ID, patient demographics, previous X-ray images, previous reports, new X-ray, new report.
- Request card medium contains: NHS ID, patient demographics, type of request.
- Cassette contains: patient identification details (vary among different radiologists) such as: NHS or unit ID, name, surname, date of birth, test report. (A cassette usually contains multiple dictations from numerous patients.)
- Report has: NHS *and* unit ID, patient name, surname, date of birth, radiograph findings.
- Radiology database record: patient demographic information, address, telephone, GP details, next of kin, etc.
- GP patient folder has all the details of the patient since he first registered with the GP.
- *Note:* Each patient has a unique NHS ID. The NHS ID is given to the patients when they are born. When a patient administers to a 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 patient ID.

Appendix B.4. Step 4: Identify the actors interacting with data and containers.

The fourth step in constructing a data journey 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. Lots of different actors can interact with the data entities we identified. However, we are identifying the ones who create, transform or consume data in order to produce some value in the scope of the journey. Actors interact with the data stored in a container. They do not interact with moving data.

Actors interacting with data stored in the container in the case study:

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

Appendix B.5. Step 5: Draw the data journey diagram

1. We start by creating the containers we identified in section Appendix B.2. We make a cube for each physical container and a cylinder for each electronic container and we name them accordingly.

2. We then connect the containers with a full arrow starting from a source container going to a target container based on the movement of data described in the process. Each arrow is a journey leg and represents the movement of data from a source to a target container. We then number the arrows in order of happening in the process for future reference.

3. For each journey leg we identify the data entities being moved and create a rectangle for data moved in physical form and a rectangle with a folded angle for electronic data based on the medium identified in section Appendix B.3. Each of the media symbols has the name of the media used (like letter, cassette, packet, etc.), and the data entities that contain (like patient ID, name, X-ray image, dictation, etc.).

4. We then put in the actors who create some value in the journey like creating a data entity, transforming it, or using it. We create an actor UML symbol for each actor and we connect it with a plain arrow to the data container which interacts to create value.

Appendix C. Other boundary diagrams

Figure C.8 shows the data container boundary. Both figures note journey legs crossing boundaries with a red warning sign. Figure C.9 shows the actors interacting with the containers to create, consume or transform data in order to produce some value.

A warning sign exists when the source container of a journey leg belongs to a different boundary than the target container of the leg. We hypothesise that each journey leg with a warning sign can impose costs to the journey of data.

References

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

[2] 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.

[3] 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.

[4] M. Y. Becker, Information governance in nhs's npfit: A case for policy specification, International Journal of Medical Informatics 76 (5) (2007) 432–437.

[5] U. National Audit Office, E-borders and successor programmes, Tech. rep., Home Office, UK (2015).

[6] P. P.-S. Chen, The entity-relationship modeltoward a unified view of data, ACM Transactions on Database Systems (TODS) 1 (1) (1976) 9–36.

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

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

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

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] I. Eleftheriou, S. Embury, A. Brass, Data journey modelling: Predicting risk for IT developments, in: The Practice of Enterprise Modeling, 2016, Springer, 2016.

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

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

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

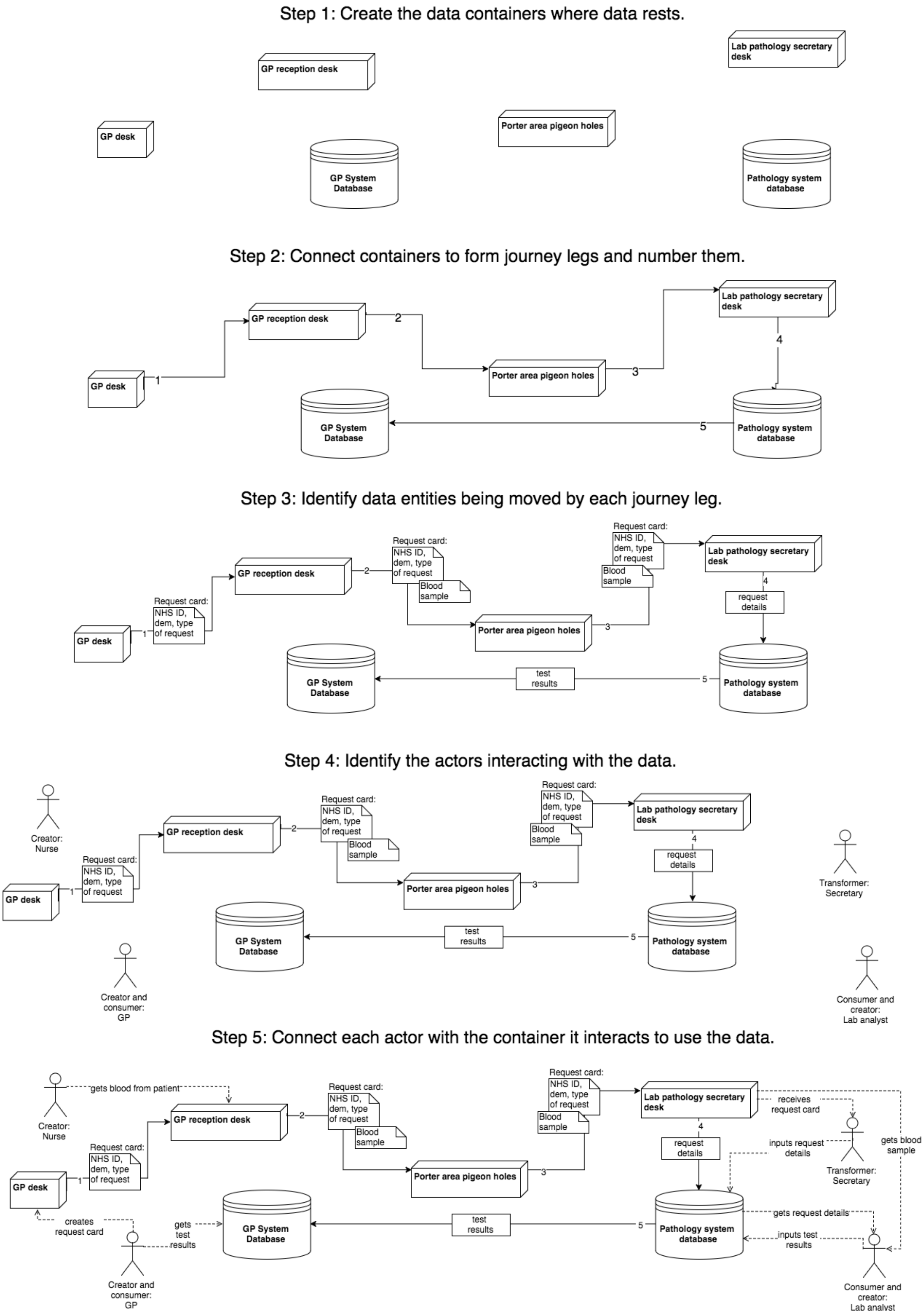

Figure 2: An example illustrating the steps to construct a data journey.

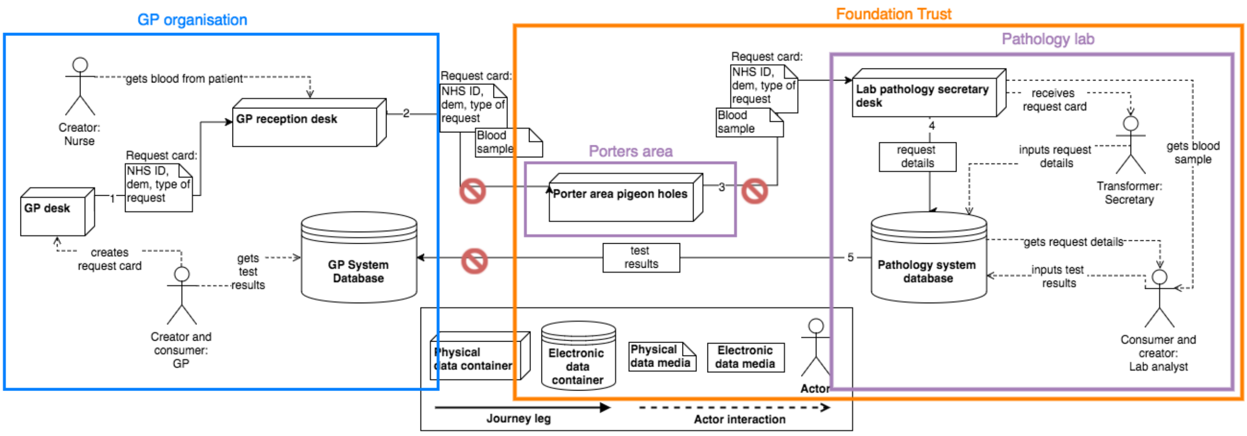

Figure 3: Organisational boundary on the GP pathology data journey.

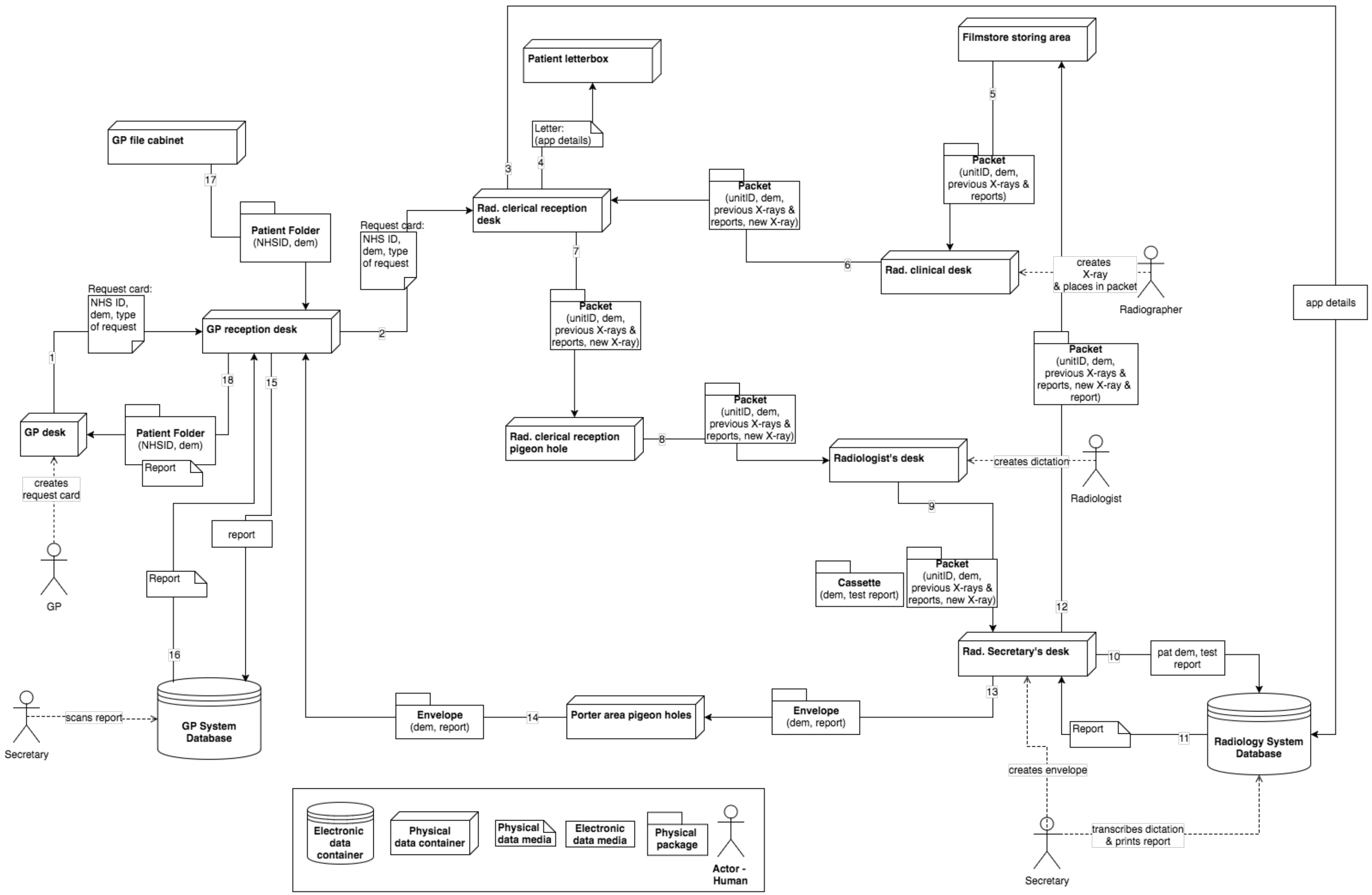

Figure 4: The journey of the data needed by a GP when a patient has a fracture at the old system.

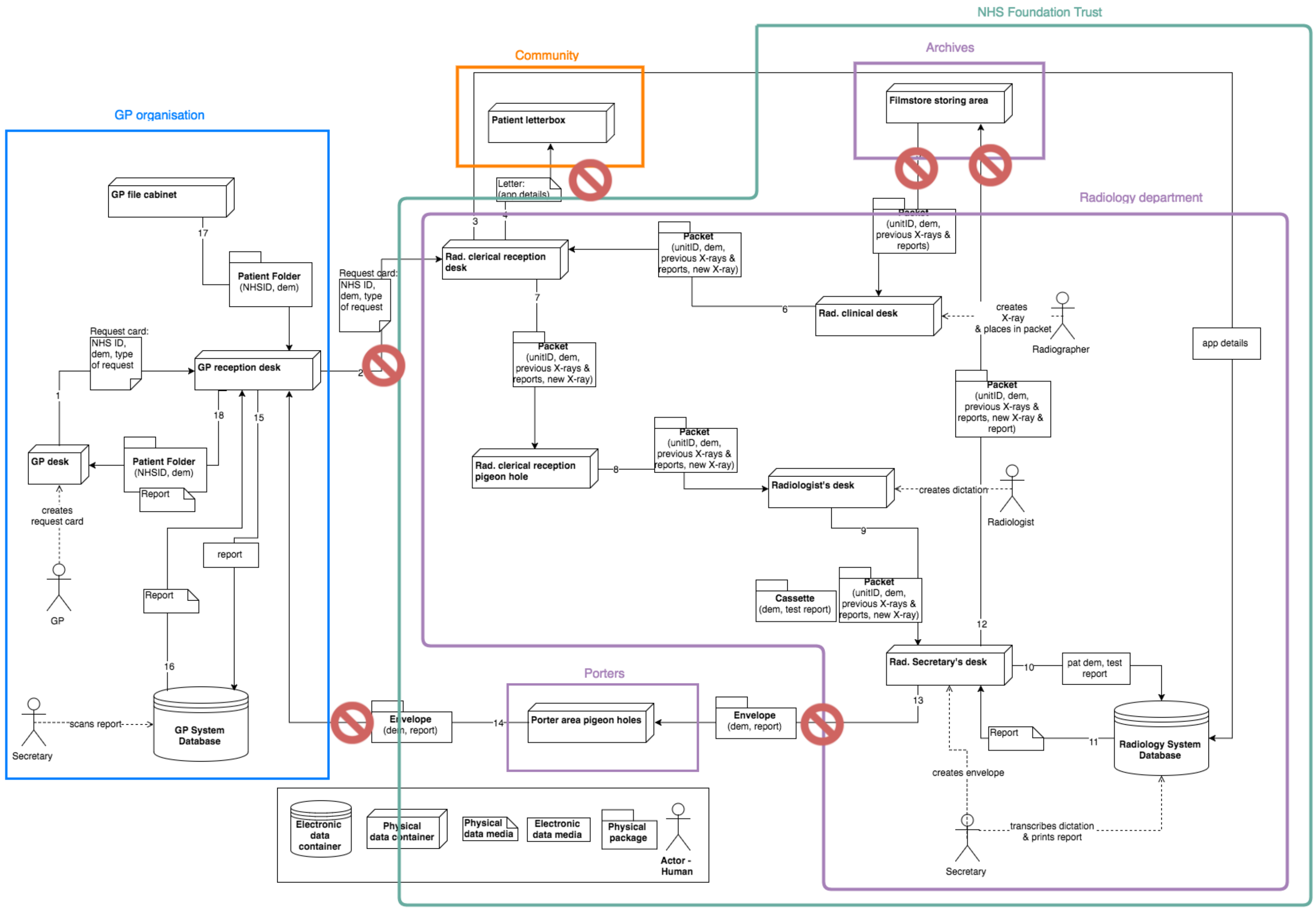

Figure 5: Data journey model of the old system with organisational structures and barriers identified.

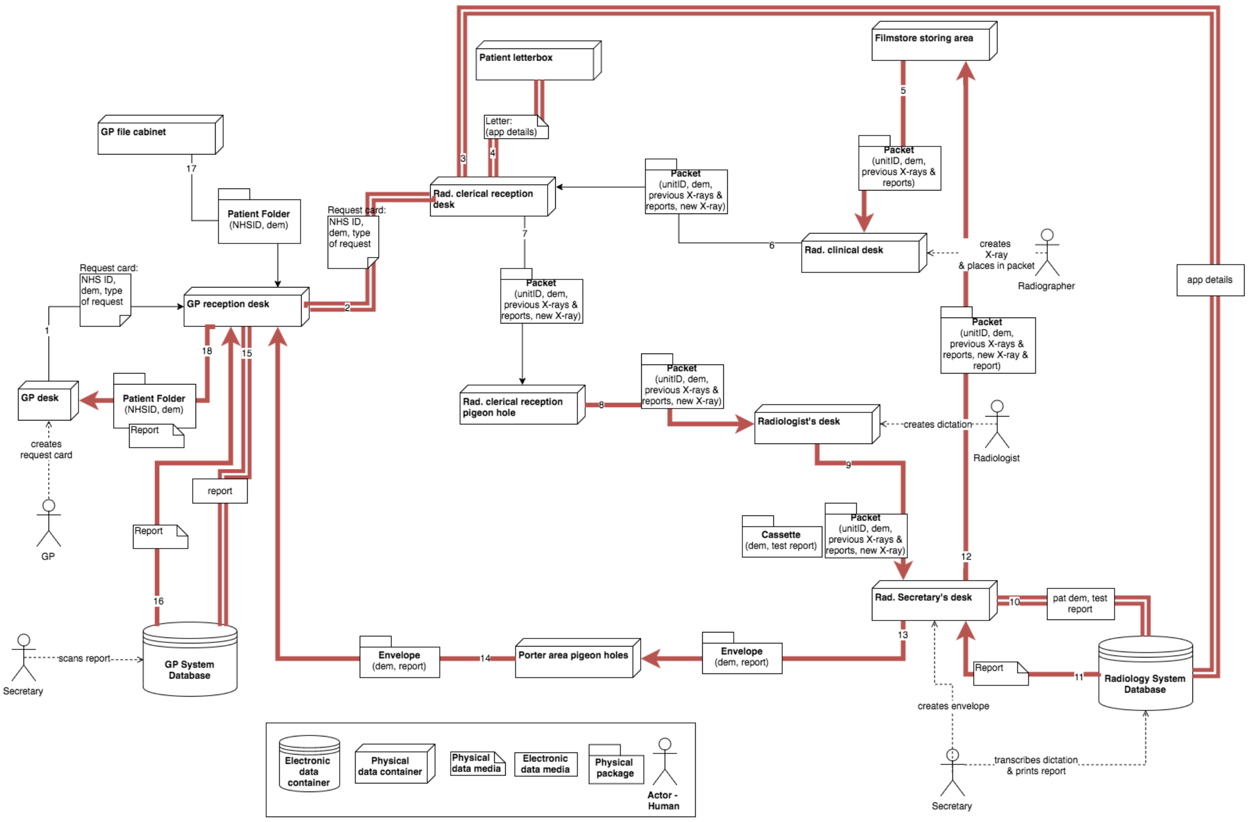

Figure 6: Heat map of the organisational, data container and actor's value boundary layers on the old system's data journey model.

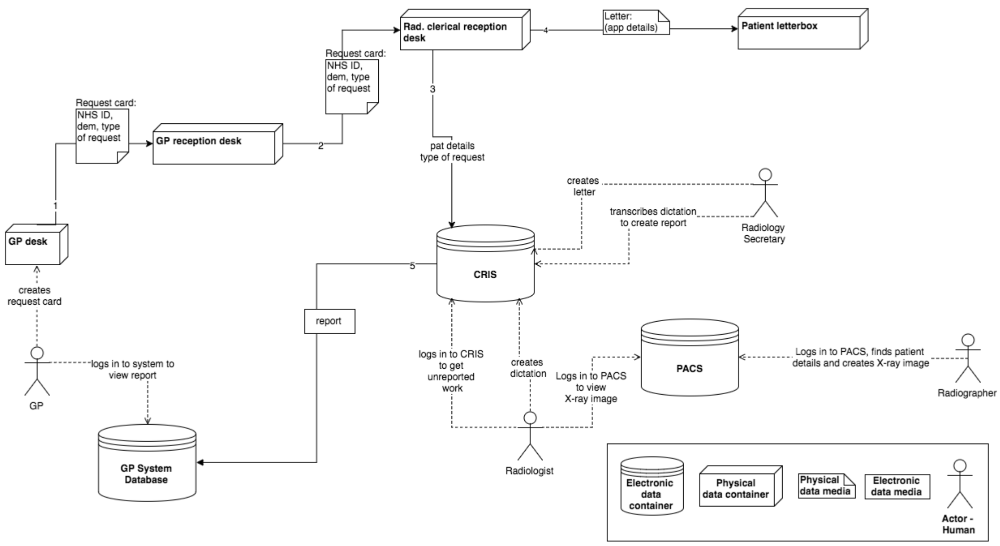

Figure 7: Data journey model of the new system.

Old Journey Leg No that changed	What changed?	New Data journey leg
1	Same data flow.	1
2	Same data flow.	2
3	Different target container, the radiology system is replaced by CRIS system.	3
4	Same data flow.	4
5	Data flow removed by replacing physical packet with electronic data saved in PACS.	-
6	Data flow removed by replacing physical packet with electronic data saved in PACS. Radiographer creates electronic X-ray image in PACS.	-
7	Data flow removed by replacing physical packet with electronic data saved in PACS.	-
8	Data flow removed by replacing physical packet with electronic data saved in PACS.	-
9	Data flow removed by replacing physical cassette with electronic data saved in CRIS. Radiologist accesses patient details through PACS, and dictates report in CRIS.	-
10	Data flow removed by replacing physical cassette with electronic data saved in CRIS. Secretary accesses dictation and transcribes report in CRIS.	-
11	Data flow removed. The report is not printed, as will be electronically sent to the GP.	-
12	Data flow removed by replacing physical packet with electronic data saved in PACS.	-
13	Data flow removed by replacing physical letter with electronic report sent directly to the GP system.	5
14	Data flow removed by replacing physical letter with electronic report sent directly to the GP system.	5
15	Data flow removed. GP accesses report directly from the GP system.	-
16	Data flow removed. GP accesses report directly from the GP system.	-
17		-
18	Data flow removed. GP accesses report directly from the GP system.	-

Table 2: Changes made to the old system by human experts and the respected journey leg number in the new model.

Success Criteria	Expected outcomes	Actual outcomes
Time	1 working day: 5 to 7 hours of clinicians time	3 meetings with P and R of approximately one hour each, and 2 meetings with R. Total of 8 hours
Effort	1 working day: 5 to 7 hours for creating the model and predicting places of high costs	??? a week or more ???
Accuracy	at least 50% of the predictions to be True Positive	8 out of the 16 journey legs that were predicted by the model to be costly, were changed in the new system. Hence, 50% of our predictions are True Positives.

Table 4: Evaluation results

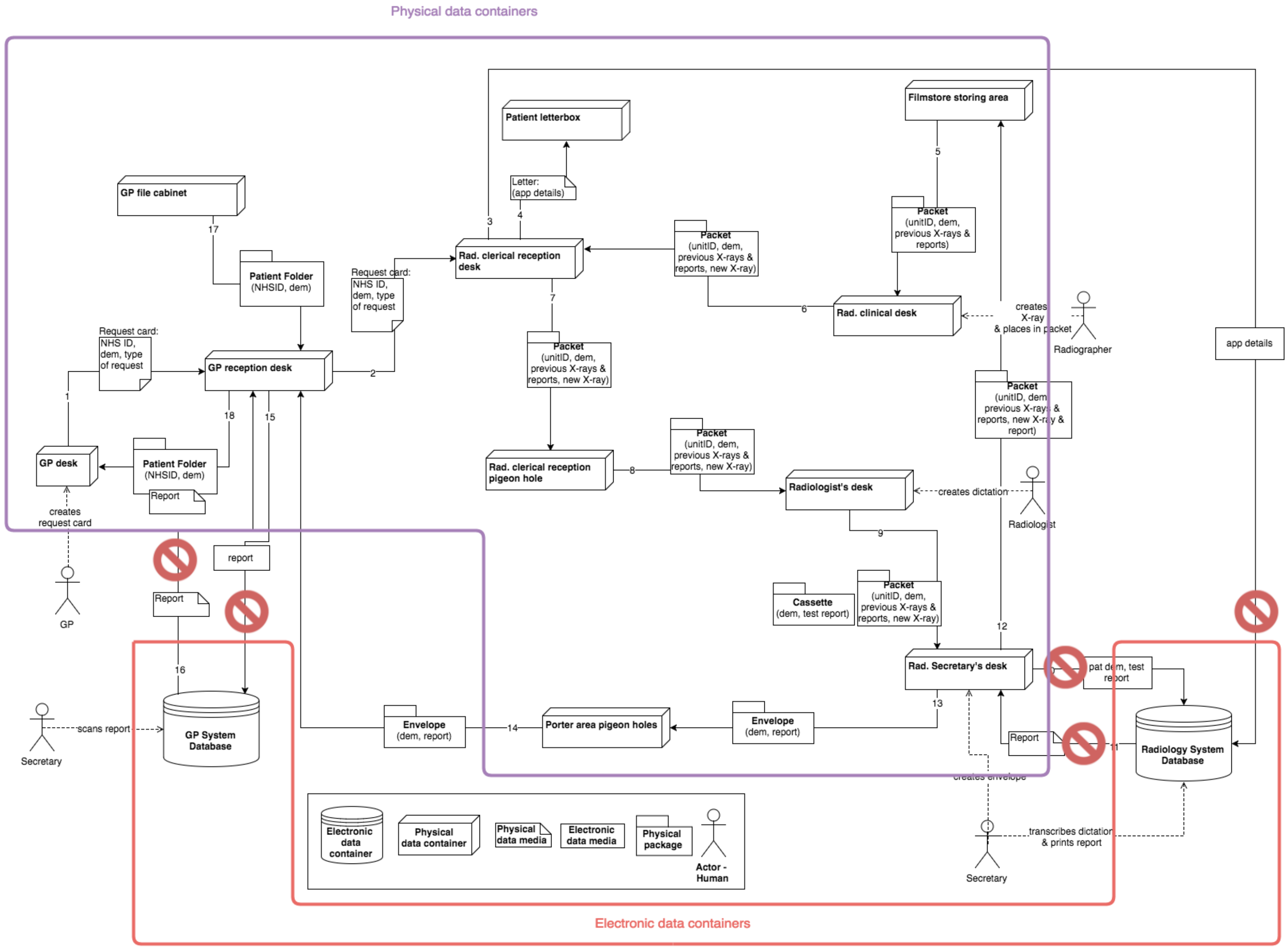

Figure C.8: Data journey model of the old system with data containers boundary and barriers identified.

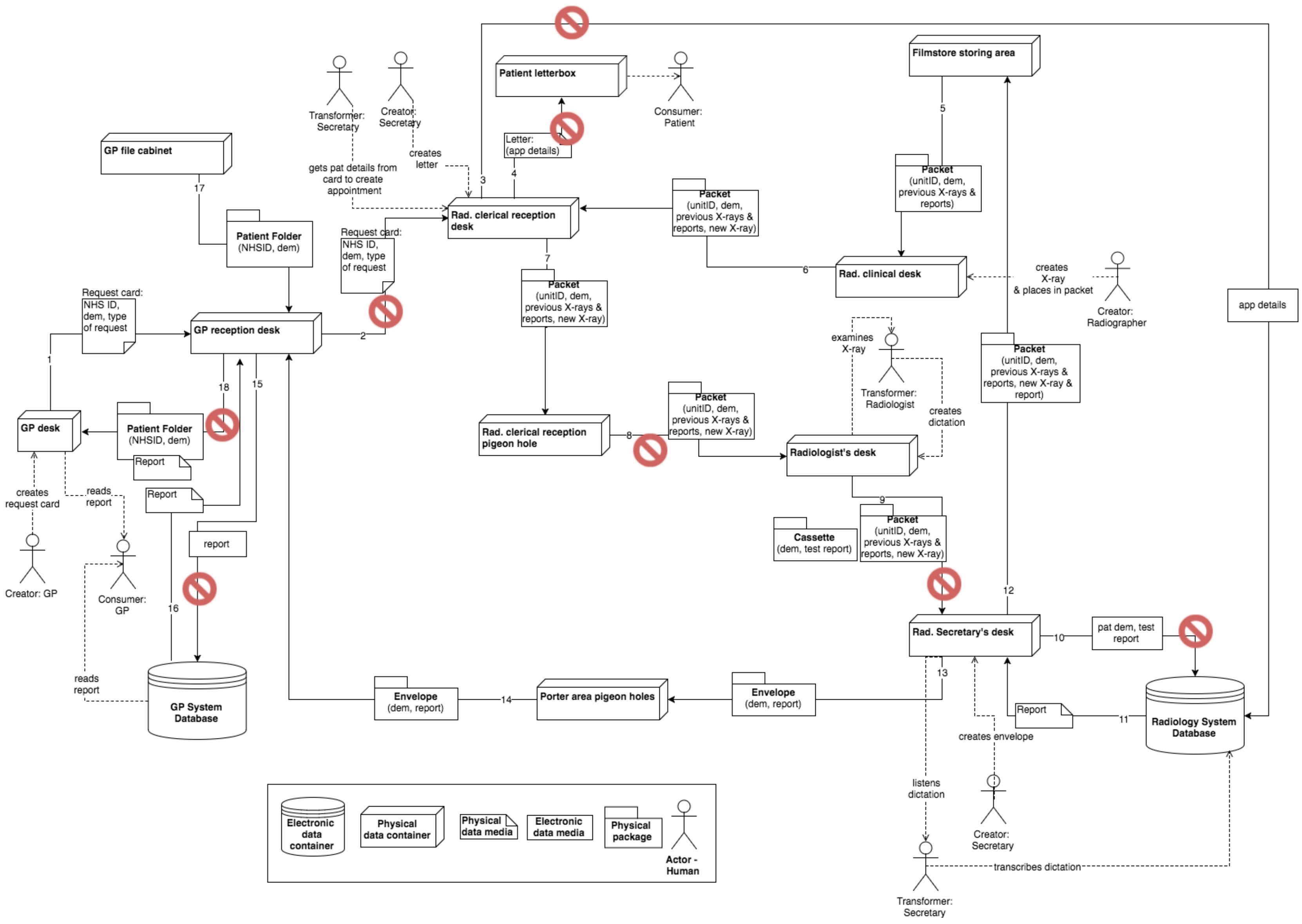

Figure C.9: Data journey model of the old system with actors' value.