UNIVERSITY OF LEEDS

Independent Repository of gambling industry data – a scoping study

Report prepared for GambleAware on behalf of the University of Leeds.

Dr Nik Lomax, Co-Director of the Consumer Data Research Centre

29 August 2019

GambleAware

Executive Summary

The University of Leeds has been commissioned by GambleAware to undertake a scoping study to assess options and provide recommendations as to how best an independent Repository of gambling industry data should be established and maintained. During the evidence gathering phase of the work we conducted in-depth interviews with 35 experts from a number of Gambling Operators, Researchers from academic institutions, two ESRC Big Data centres, GambleAware and the Gambling Commission.

There was wide scale support from all groups for the establishment of an Independent Repository of gambling industry data. Operators see a Repository as a way of supporting research into gambling harms and providing results and methods that can be used within their own business. Researchers see a Repository as a way of providing access to data which has historically been difficult to obtain and as a way of legitimising research into gambling. The Commission see the Repository as a pillar of their new National Strategy to Reduce Gambling Harms.

There was agreement that *someone* needs to take responsibility for housing and maintaining a Repository but there was no immediate consensus on who this should be. When pressed, there was general agreement that a university would be a suitable institution, given their independence from the gambling industry and obvious link with the Researchers who would want to use the data held in the Repository.

Three key barriers were identified which will need to be dealt with in order to successfully establish a Repository: (1) legal, data protection and ownership of data; (2) the consistency of terms and datasets across different Operators; and (3) identifying what data would actually be required. Pragmatic solutions to the first set of issues can be dealt with by adopting best practice from existing Data Centres. The second and third issues around consistency and data requirements will require collaboration between stakeholders to agree on suitable terms of reference.

Building trust with data providers and data users is key to the success of a Repository. We suggest an approach we have termed the ‘Ladder of Engagement’ where a small number of highly engaged data partners occupy the top rung, providing datasets which can be shared. Lower rungs are occupied by other partners who engage in specific, smaller scale activity. Within the framework of a Repository, partners can be moved up the rungs of the ladder as relationships are developed.

Data acquisition processes and infrastructure for secure housing of data need to be robust. We set out key processes used by the Consumer Data Research Centre in this report and we highlight that having adequate staffing for the curation of data and for acting as a trusted third party to Operators and Researchers is essential for the success of a Repository.

Set up and maintenance costs of a Repository could be substantial. We suggest that aligning the Repository with an existing Data Centre could help reduce these costs at the outset and also have the benefit of offering existing established processes and governance structures which would otherwise take time to set up.

In summary, our headline recommendations are:

1. Start with a smaller number of highly engaged stakeholders who can help build the Repository. Be prepared to engage with Operators at different levels rather than expect all to be able to share data immediately.

2. As far as possible, work collaboratively to agree on definitions and data requirements.

3. Establish strong legal agreements between Operators and the host of the Repository for a sustainable relationship.

4. Establish the Repository within a university, or at least ensure that a university is closely aligned with the work.

5. Align with an existing Data Centre to reduce costs, draw on expertise and borrow established governance structures.

6. Ensure that a Repository has adequate staff who can represent the Repository as a trusted third party.

7. Player data needs to be linked across different Operators in order to provide in-depth research on the harms of gambling.

Specifically, our short term recommendations are to:

1. Work with a group of highly engaged Operators, Researchers and an existing Research Centre to define research questions, establish processes and iron out problems.

2. Scale up the Repository to attain critical mass and become self-sufficient.

3. Obtain immediate funding from Operators and seek to leverage Research Council funding.

Contents

1 Introduction and scope of report

The University of Leeds has been commissioned by GambleAware to undertake a scoping study to assess options and provide recommendations as to how best an independent Repository of gambling industry data could be established and maintained – to provide maximum access to data for Researchers, benefits to policy-makers and industry, with the appropriate safeguards for consumers and gambling Operators.

The gambling industry generates significant amounts of data, which can be analysed to gain insight into how different groups of people gamble, on different products and in different gambling environments. This data are important for all stakeholders in order for them to understand whether some gambling products or environments are more harmful than others. This can inform interventions and policy to minimise gambling-related harms and promote safer gambling practices. In addition, the data are valuable for Researchers in general to investigate a range of other research questions including those related to spending, lifestyles and attitudes to risk.

This report sets out the findings of the scoping study. In collecting the evidence presented in this report, we have engaged with Gambling Operators, Researchers who use gambling industry data, UK based academic Data Centres, the Gambling Commission and Gamble Aware.

In preparing this report we have drawn heavily on our experience of running the Consumer Data Research Centre (CDRC), a Data Centre based at the University of Leeds, which to date has received over £6.5 million in support from the Economic and Social Research Council (ESRC) to engage with commercial data providers, license these data for use and undertake a substantial programme of research in areas as diverse as business, psychology, sociology and social science. We have also engaged extensively with the Urban Big Data Centre[1], another ESRC investment based at the University of Glasgow, with a focus on using urban data to address global city challenges, as well as the wider research community to better understand requirements and best practice in setting up and maintaining a Repository.

2 Approach taken and structure of report

In preparing this report we undertook face-to-face in-depth interviews with Operators, Researchers and other stakeholders. This included interviews with 35 experts from:

- Four large gambling Operators
- GambleAware
- The Gambling Commission
- NatCen Social Research (referred to as Researchers in the report)
- The Urban Big Data Centre
- Consumer Data Research Centre
- Academic institutions (referred to as Researchers in the report)

[1] https://www.ubdc.ac.uk/

These interviews were broadly structured around the requirements set out in the GambleAware project brief. This report draws on those key themes, where possible leading with the issues identified during the interviews. We offer solutions and additional information based on our experience from running the CDRC and other desk-based research. In order to preserve the privacy of those who took part in the interview process, we have anonymised individual responses.

The remainder of the report is structured as follows:
In Section 3 we summarise the general views provided by our interviewees on setting up a Repository of gambling industry data; Section 4 sets out some options for housing and maintaining such a Repository; in Section 5 we summarise some of the barriers, and offer some solutions for setting up a Repository; Section 6 sets out how we build trust with data partners and research users who engage with the CDRC; in Section 7 we discuss data acquisition processes and infrastructure requirements; Section 8 lays out some options for funding and potential costs of setting up a Repository; finally in Section 9 we offer some conclusions and recommendations.

3 General views on a Repository of gambling industry data

By way of framing this report, this section outlines the broadly supportive view expressed by those we engaged for setting up a national Repository of gambling industry data. It outlines some of the motivations of Operators, Researchers and the Gambling Commission for establishing such a Repository.

Researchers

We interviewed a number of academics who use gambling industry data in their research. This group immediately drew attention to the fact that, compared to other areas of study, there are a small number of academics who are active in gambling and betting research. The main reasons for this were attributed to:

1. The lack of funding within the UK; and
2. the lack availability of easily accessible data.

On the first of these points, the attractiveness of a Repository was directly connected to the availability of funds to make it a success, summarised by one researcher:

> *“There is no point in creating this massive data Repository if there is no funding resource to be able to support it or create a community of people to make use of it”*

The interviewees agreed that if a UK National Repository could provide quick and easy access to data then betting and gambling would become a more attractive research area for UK academics. However, to reach this point it was anticipated that much work was needed to make the Operators’ data compatible. Researchers reported that any data collected needs to provide detailed insight in to consumer behaviours. Data would need to be more than top level and aggregated in order to be of interest. The Repository has to be able to provide deep and rich data as this will enable Researchers to investigate the questions they wish to study. There was particular interest in the ability to link information on users across different industry Operators, particularly where the focus of research was on the harms of gambling, where customers are likely to have accounts with multiple providers. We discuss the issue of linkage in Section 5.

It was also stated that there was often an uncomfortable view within academia on researching the betting and gambling industry due to perceived issues of ethics and the independence of Researchers. We pick up on these ethical issues in Section 5. It was believed the creation of a National Repository could address this unease, though how such a Repository was established and funded would determine whether it would alter the stigma of researching in this sector.

Operators

Staff at a number of large Operators were interviewed as part of the evidence gathering for this report. All the Operators interviewed were committed to proactively assisting their customers who identified themselves or who the Operator identified as having problems with their gambling habits. Some had research programmes and tools in place to help customers and appeared keen to develop and move the effectiveness of these activities forward. In addition to their own efforts in these areas, it was believed an increased and wider base of Researchers undertaking studies would be of benefit to them and the Industry as whole.

Operators recognise that research carried out by third parties especially by Academics will be undertaken with rigour and using valid techniques. This will give the dual benefit of offering results and methods which can be incorporated into their business practices and allowing the outside world to see that they are supporting independent research on gambling harms. One Operator summarised this opening up of data and collaboration with Researchers:

> *"If you encourage everyone to contribute and they get something back then you will get more information and better quality data and so the industry will improve"*

The Operators believed the concept of the Repository should not be that they stipulate the research undertaken but that research findings should be fed back into the Repository and for them to have access to the reports in order to use the findings to adapt their customer practices where appropriate.

> *"The research is not only to advance the learning of Academics, it is also to inform the Industry on potential enhancements to its existing approaches so everyone has a vested interest in the outcomes"*

One Operator proposed the data in the Repository could be used to develop a series of Industry benchmarks against which individual Operators could then judge their own performance.

“Comparative information is always valuable and drives good behaviour”

A Repository of data was seen a valuable instrument to achieve this increase in research activity. However, the opening remarks of the Operators always focused on the practicalities of establishing a Repository that could be useful and useable for research purposes. These comments referred to the need to bring together the different terms, descriptors and data formats used within the Industry.

This perceived lack of consistency in definitions and data held between Operators is seen as the major barrier to the extent that one of the Operators suggested that rather than investing time and resource in establishing a Repository, data might be provided in response to specific research requests. It was thought this approach would avoid lengthy and potentially intractable discussions between the various parties about definitions and processes. This key issue of definitions is discussed in more detail in Section 5.

The Gambling Commission

The concept of a Repository has been on the Gambling Commission’s agenda for several years and development forms a fundamental pillar of their new ‘National Strategy to Reduce Gambling Harms’[2].

All interviewees from the Gambling Commission believed a National Repository would help with decision-making in this sector by:

- leading to an increase in the amount of evidence-based research which is undertaken; and
- creating a wider pool of Researchers which would result in new thinking and additional skills being applied to research in gambling.

Colleagues acknowledged that there could be a spectrum of approaches to acquiring data from Operators. These ranged from “a gradual voluntary” approach to a more “hard line” perspective in which the building and maintenance of this National Repository and the internal costs of putting data into a standard order should all be seen by the Operators as a cost of business.

All the interviewees put forward the view that the introduction of the National Repository should be a phased process. This was seen to be important for the following reasons:

- to build credibility; and

[2] https://www.reducinggamblingharms.org/

- to get it right for the long-term.

Ultimately however, once fully established it was made clear that all Operators should have to participate. It is at this point that the Commission could back this up with licensing requirements if necessary. The reality is that sometimes the bigger, more willing operators take part, but the smaller ones or those who are less willing to engage don't get involved until there is a formal requirement.

It was identified that the Repository's management and governance should be seen to be independent of Operators. At the same time, a counterview was expressed that the creation of another body within the Industry was unnecessary and may cause overlap and thereby muddle decision making and action. It is clear that any governance processes established as part of a Repository should not be unnecessarily restrictive or burdensome on the Industry.

Similar to the view expressed by Operators, it was appreciated that there was considerable variation in how data is held, formatted and defined across the sector. One solution offered for tackling this inconsistency was to have Operators help produce and then conform to a single set of definitions and data rules. This might stem from Researchers specifying the data needed.

Summary

- There was wide ranging support for the establishment of a Repository of gambling industry data from Operators, Researchers and the Commission.
- A key issue identified by both Operators and the Commission is that of consistency in definitions and the types of data that could be supplied. We discuss in Sections 5 and 6 some strategies for collaboration between data users and data providers which should help tackle these issues. We also suggest that these specific data issues should be addressed in the *Patterns of Play* work being undertaken by NatCen[3], which is part of the Commission's wider research programme.
- The establishment of a Repository was not seen as a quick fix, but rather something that would require long-term investment and cooperation between different parties.
- Ultimately all Operators would be expected to contribute once the Repository is fully established.

4 Options for housing and maintaining a Repository

There was a consensus amongst Operators and Researchers that there should be a person or group in charge of the Repository, summarised by this statement:

> *"Someone needs to own it because someone needs to be accountable and responsible for the care of that data."*

[3] NatCen Social Research are undertaking a wide programme of work which includes the *Patterns of Play* project which is addressing the type of data which could be made available from Operators.

However, the question of where a Repository should be housed and who should run it was one on which there was the least clarity and conformity of opinion from the researcher and Operator groups interviewed. Distinction was made between the holding of the data and the management of the data with divergence of opinion: some respondents saw these as being handled in the same place and some seeing them as possibly separated. There was a view from the majority of respondents that the management and responsibility for a National Repository needed to be independent and should have its own rules of governance and be separate from any existing bodies. However, some Operators expressed the view that the industry did not need another governance body.

When prompted (from a range of options including the Commission, a University, a business hosting e.g. Microsoft/Google), Universities or a group of Universities were generally seen as bodies with the necessary independence, credibility and appropriate structures to hold and manage such a Repository.

"Universities have the right intellectual rigour to challenge the data"

We lay out some examples below.

University or Academic Institute repositories

In this report we have drawn on our experience running the CDRC. However there are other academic institutes which store, process and make available datasets from external providers. The CDRC was funded under the ESRC Big Data Network and established to make data routinely collected by businesses and government departments accessible for research purposes and in the interest of the public good. The network also funds the Urban Big Data Centre (UBDC) based at the University of Glasgow and the Business and Local Government Centre (BLG) based at the University of Essex. Many processes within UBDC and BLG are similar to those we have implemented for the CDRC and as part of the evidence gathering for this report we have worked with UBDC to understand their processes in greater detail. Where appropriate we have integrated these findings in the report. Similar in set up and scope was the Administrative Data Research Network (ADRN) based at several sites around the UK, set up to provide an infrastructure that allowed social Researchers to use administrative data in a safe setting. The ADRN funding came to an end in July 2018.

Another example is the UK Data Service (UKDS), also an ESRC funded resource, based at the University of Essex (and supported by other Universities: Manchester, Edinburgh, Southampton). The data stored by UKDS includes major UK government-sponsored surveys, cross-national surveys, longitudinal studies, UK census data and qualitative data. UKDS is well established and in many ways operates in a similar way to the Big Data Network centres in terms of processes for acquiring and making data available.

Other, open-access repositories exist for academics and other providers to deposit their data, although these are less suitable for the purpose of setting up a Repository because there is little infrastructure governing access to data. Some examples include FigShare, Zendo and Dryad. We are not advocating these types of service because of the need to control access to datasets deposited by Operators.

What the Big Data Network and other Data Centre investments detailed here demonstrate is that there is appetite in UK academia to fund projects which make datasets available to Researchers.

Summary

- The groups interviewed were not especially clear on who should be responsible for housing and maintaining a Repository.
- When pressed, a university was identified as the most obvious place to house such a Repository, given the connection to those undertaking the research.
- Examples from the UK academic sector demonstrate that there is clear demand for Data Centres which connect data providers with data users.
- There are a number of successful examples of Data Centres which have established procedures and mechanisms for storing and sharing data.

5 Barriers and solutions to setting up a national Repository

Setting up a National Repository requires trust and investment from all stakeholders, not least from the Operators who will be supplying data. This section covers some of the key issues raised by Operators and Researchers during evidence-gathering which will need to be addressed in order to make a National Repository a success. These issues are broadly centered on: (1) data protection, legal issues and ownership of data; (2) agreeing on definitions and consistency for datasets; and (3) difficulties in specifying what kind of data would need to be supplied to a Repository. Due to the complexity of these issues this section relies heavily on our first-hand experience within the CDRC (with additional evidence from UBDC) to exemplify the barriers and solutions which would be faced more generally by anyone attempting to set up and run the Repository.

Legal issues, data protection and data ownership

During the interviews, legal, data protection (including GDPR) and ownership were all mentioned as issues that would need to be addressed if Operators were to contribute data to a Repository. However, it seemed to be assumed that these matters could be resolved by reference to professional advice or other existing models in other sectors. This is a bold assertion because these are some of the most complex issues to deal with when handling data from multiple organizations. In the sections below, we spell out some of the ways in which we handle these issues within the CDRC, with additional examples drawn from the Urban Big Data Centre (UBDC). Like the CDRC, UBDC has, until recently, provided a national data service offering access to open, safeguarded and controlled data assets.

Legal Agreements Underpinning Data Service

Securing legal agreements is essential when building a relationship with a data provider. The CDRC has several legal agreements which underpin the data service. The Centre has **Data Licence Agreements (DLAs)** in place with data providers which allow for the sharing of data with Researchers and other approved users of the national data service. Terms and conditions for data sharing are agreed in a standard contract between the University and the data provider (wherever possible) but recognising unique circumstances in many cases. We can make this agreement template available on request. Our standard contract enables the University to grant sub-licences called **User Agreements** to users to enable their access to data for research purposes, while **Data Sharing Agreements** allow the sharing of data across multiple sites within the data service.

Data Licence Agreements (DLA) contain key clauses, among other things, around use of the data, ownership of intellectual property and consultation and approval of publications (academic papers, reports, etc.) arising from research using the datasets. They also make clear the service tier (see next section) through which the data are to be made available to users, and a number of conditions such as whether the data are to be archived or deleted on expiry of the agreement, and whether the University will be allowed to name the data provider as a partner and use their logo.

UBDC licensing

UBDC has no standard DLA template. While this may mean that they are able to sign off more quickly by agreeing to terms and clauses that are routinely used by the partner, it does potentially pose issues around implementation. DLAs for safeguarded data vary considerably, resulting in widely differing stipulations depending on the data provider and the specific asset. It also results in their being no single user journey for safeguarded data access. For their controlled service, DLAs are negotiated as part of the access process. This may result in lengthy processing timeframes and also presents a degree of risk and uncertainty for the researcher.

User Agreements are issued to all users who have been approved to access either safeguarded data and/or controlled data defined in the next section). Such agreements are made between the user and the university. User Agreements set out the terms and conditions of use of the data as well as the penalties that may be imposed for breaches of security or confidentiality, or any other breach of the agreement. Signed schedules appended to the User Agreement set out all of the relevant stipulations relating to use of the data, including those determined by the Data Partner, for example, around training requirements for users, timelines for data use and deletion, as well as conditions around data partner anonymity. User Agreements must be signed off by all Parties prior to access being granted to data.

The third agreement used by CDRC is the **Data Sharing Agreement**. Because the CDRC is a national service, led by the University of Leeds and University College London (UCL) (with collaborators at Liverpool and Oxford) this agreement allows for data acquired by any one institution to be shared with the others for the CDRC service. Our standard Data Licence Agreements make explicit reference to this Data Sharing Agreement. Therefore in agreeing to share data with the University of Leeds, data providers also consent to the sharing of their data across partner institutions. This might be important if a Repository of gambling industry data were a joint venture between a number of partners (e.g. a University and Gamble Aware) with access provided in/by those partners.

Defining Data Service Tiers

The CDRC provides access to data via three service levels: open, safeguarded and secure. Broadly speaking, these levels correspond to the service levels described in the UK Data Service's three tier access policy[4]. Defining data under these service levels is useful as it allows us to collect data for different purposes and make data available to a wide range of potential users. The burden (cost and time) of administering open access data is less than that of secure data. In the context of a Repository of gambling industry data, offering different access tiers might be attractive as it would allow a wider range of researchers to engage with the service.

CDRC Open Service: Open data can be accessed via basic registration and download from our data portal. These data are freely available to all, for any purpose, and include aggregated data and derivative products produced by the CDRC.

CDRC Safeguarded Service: Safeguarded data are not considered 'personally-identifiable' or otherwise sensitive, but are data to which access is restricted due to license conditions. This might include data from retail companies on store turnover. Access to such data is subject to application and approval and via secure download.

CDRC Secure Service: Controlled data which need to be held under the most secure conditions are provided via the CDRC's secure service. These data include data that are 'personally identifiable' and therefore subject to Data Protection legislation or are commercially sensitive. This service requires that individuals gain project application approval and visit one of our secure facilities at either the University of Leeds, University College London or University of Liverpool to access and use the data.

[4] https://www.ukdataservice.ac.uk/get-data/data-access-policy.aspx

UBDC Controlled Data Service (currently suspended pending decisions on funding)
UBDC provides a Controlled Data Service, but does not maintain a collection of controlled datasets. Rather, the Centre indicates to prospective users which data may be available and which organisations it has ‘in principle’ data sharing agreements with. In this way, it brokers access with data owners to enable researchers to work on specific projects with specific datasets. To access controlled data, users need to go through a formal approvals process which begins with the submission of an initial online application. If the initial application fits the remit of the Centre and appears feasible, then the user is assigned a member of staff (typically a data scientist) who will advise on the development of a full proposal. Researchers are also required to obtain ethical approval for the project. At this stage, preliminary discussions are held with relevant data owners to secure ‘in principle’ agreement for data sharing before proposals are submitted for independent assessment by the Research Approvals Committee (RAC). In addition to RAC approval, data owners also need to give their approval to the project.

Once both the RAC and the data owner approve a project, the researcher’s institution and the data owner need to sign a data sharing agreement. A privacy impact assessment (PIA) may also need to be completed in order to outline the potential risks to privacy and how these risks will be mitigated. Data sharing agreements may effectively make the researcher’s institution a data processor for the data under the Data Protection Act (DPA) 2018. UBDC’s controlled data service is facilitated via a respected provider, eDRIS (electronic Data Research and Innovation Service), part of NHS National Services Scotland, who provide a highly secure environment. As a result, the researcher’s institution and eDRIS also need to enter a data processing agreement which permits the latter to receive data on behalf of the researcher. Finally, the researcher will be required to sign End User Licences with the data owner, with UBDC and also a User Agreement with eDRIS.

Consent and Anonymisation

EU General Data Protection Regulation (GDPR) is the overarching legislation covering the collection and use of personal information. If any **identifiable** information (name, address, etc.) is to be deposited in and processed by the Repository then the information provider – in this case the gambling Operator – will need to evidence that consent has been captured from each data subject (i.e. the customer or player) to carry out research. GDPR expects this to be transparent and a general consent to the use of data for research purposes might not be enough.

It would be useful in future work to understand how and if Operators obtain consent from their customers for data to be passed on for research/ other purposes, how this consent is obtained and what the extent of the consent is. Some research activities, for example linking datasets, would require the explicit consent of the customer, which is why general consent may not be sufficient.

If one can demonstrate clear physical and legal separation between Researchers using the data and the link back to the individual, and there is no realistic change of disclosure from the data, then one should be able to declare the data as anonymous (this used to be known as ‘anonymous in context’ which was a much more helpful term), and thus outside the purview of GDPR. However, this has yet to be tested in law, and is the subject of much anxiety amongst Researchers – disclosure is very difficult to imagine, but not impossible.

In principle you are expected to justify use of all data, and if you can, to reduce risk of disclosure. One example is that data fields could be grouped where appropriate, e.g. transforming single year of age to age bands.

Linking of datasets

From conversations we have had with Researchers, there is a preference for a Repository to be developed with the ability to link individual gamblers across multiple Operators. This might allow for a wider market view of player activity, for example when identifying where users have multiple accounts.

In theory the data from different Operators could be linked, given that each Operator holds information which is unique to a player. However undertaking this kind of linkage would require specific consent and the mechanisms for undertaking this linkage will need to be thought about very carefully. It is often explicitly stated in data access agreements (e.g. with much of the data held by the UK Data Service) that data will **not** be linked by Researchers, so this is an activity which would likely need to be undertaken in-house, within the Repository. One mechanism for doing this would be through pseudonymisation of sensitive data fields.

Pseudonymisation

Risks under GDPR and the Data Protection Act (DPA) 2018 lie in the identification of individuals. Identification could come from individual fields (e.g. name) or a combination of fields (e.g. postcode and date of birth). In the context of the draft NatCen *Patterns of Play* specification of potential datasets shared with us, player ID will be linkable back to an individual person by the Operator. If the link is not available to the researcher (or the Repository) then the data might be said to be pseudonymised. GDPR still applies to pseudonymised data, but there should be some, probably significant, risk reduction.

Pseudonymisation is described in the GDPR legislation as *“the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information.”* Such additional information should be kept separately.

One reason for undertaking pseudonymisation is that datasets held on servers often need to be linked to provide utility, and this linkage requires some identifiable (i.e. disclosive) attributes within each dataset to ensure the linking is done correctly. One way of doing this is by removing identifying attributes (for example name, address or location) through anonymization but this poses problems for the linkage of data held in a distributed environment. The Information Commissioner’s Office code of practice recognizes the value of pseudonymisation for linkage in that *“pseudonymised or de-identified data may be very valuable to Researchers because of its individual-level granularity and because* *pseudonymised records from different sources can be relatively easy to match."* So in summary, fulfilling the needs of research users in linking datasets may be possible but will require further work to ascertain the practicalities of doing so, especially on the scale required.

One way of providing this pseudonymisation is by using a technique called hashing. This involves replacing identifiable information with a pseudonymised string called the digest. The Secure Hash Algorithm version 2 (SHA-2) is the current standard, created by the US National Security Agency (NSA) and published by the National Institute of Standards and Technology. The deterministic nature of hashing means that, given the same original input value and hashing function, they will always provide the same output. This is important for linking datasets, since the same digest can be generated for multiple datasets where the original input values held within a secure research environment. In terms of data dissemination, including the hashed value as part of the release table allows users to request further information without jeopardizing the security of the dataset. It also allows for the follow-up of an individual record within a safe environment. It can be combined with anonymization techniques to provide different levels of security for datasets. We use hashing techniques to link datasets in the CDRC. In the interest of space we have not expanded on methods of pseudonymisation here, but can provide further information if required.

Linking: an example from NHS Digital

NHS Digital are working on linking patient data to other resources to provide linked datasets to support health and care delivery and research. Many of the linkage attributes are similar to those which might be available from gambling Operators (name, date of birth, postcode). NHS Digital have made their methodology available[5] (they use deterministic algorithms) but of interest here is the way in which data are stored and used. NHS Digital have set up robust procedures for safeguarding data and a process by which Researchers can apply for access and undertake bespoke linkage. In the same way in which CDRC hold and process data the key principal is that data access is secured, those applying for data are vetted and there is a focus on non-disclosure of identifiable data.

Disclosure

Disclosure relates to the risk of identification of an individual. This relates to disclosure within the raw data and disclosure from outputs of research. A robust research environment will mitigate many of these risks.

Within the raw data there may be an opportunity for a researcher to identify an individual through a single field or a combination of fields, especially if identifiable information is being stored. This risk can be mitigated by having controlled access to datasets, procedures that are well established within the Consumer Data Research Centre and more widely, for example the UK Data Service. This means that Researchers sign up to an agreement which

[5] https://digital.nhs.uk/binaries/content/assets/website-assets/services/dars/linked-datasets-in-nhs-digital-final.pdf

allows them access to data within a safe setting, e.g. a data safe haven. It is the responsibility of the researcher not to disclose identifiable information.

There is potential for disclosure from the outputs of a research project, for example where small numbers are reported or unusual characteristics or a combination of characteristics are present. These risks can be (and are routinely) mitigated by using robust disclosure control methods. These include ensuring that outputs meet a certain criteria (e.g. no counts under 10).

Commercial Sensitivity

Some issues that a Repository might face are around commercial sensitivity, as opposed to disclosure. This includes the potential for housing data that might give a competitor a commercial advantage, or that disclosure of data might cause reputational damage. Key to risk mitigation in this context is to ensure that there are appropriate controls in place when storing and accessing the data, as detailed in Section 7.

Intellectual Property

In the CDRC process, intellectual property (IP) is generally negotiated with the data provider as part of the Data Licence Agreement. The CDRC has standard clauses. Data supplied for research purposes is under licence for a specified period of time, so the Repository would not own those data.

In principle, the standard clause is that any IP rights arising from the use of data by the university shall vest in the university (in the context of this report the university would be the institution hosting a Repository). Any IP arising from the use of data by a researcher will vest solely with that user.

Ethics

One of the barriers to using gambling industry data identified by Researchers we interviewed is that the research community feels uncomfortable with the Gambling Industry and its perceived ethics.

> *"Currently the perception is that Researchers are in the Industry's pocket."*

Therefore in establishing a Repository, a way has to be found of demonstrating that the Industry is not in control of the Repository nor dictating the usage of it. Amongst the Operators, the term transparency was frequently mentioned, in terms of the use of data and governance of the Repository.

One way to deal with the ethics of storing and using data is to draw on existing processes, for example the guidance set out by University ethics committees. At Leeds, CDRC practices adhere to guidance laid out by the University Research Ethics Committee (UREC). These processes are quite lengthy to describe but are clearly documented by the University[6] and further information can be supplied. Of note, The CDRC Research Approvals Group (RAG), described later in this document, deals with the ethics of research undertaken using CDRC data. We would suggest a similar approach for a Repository of gambling industry data, to ensure ethical considerations are part of the workflow for undertaking research.

Definitions and data consistency

An issue of key concern was the bringing together of Operators data from different platforms where definitions, structures and processes were currently not compatible. Without addressing this barrier, it was felt that data would be of limited value to Researchers. There were some clear statements on this issue:

"The first challenge is around consistent definitions."

"Defining the data and defining the format of the data will be a big one."

"An absolute prerequisite for the Industry is to agree on what things mean."

When asked about the possible time scales for formulating these definitions, the interviewees were uncertain though when pressed the indication was probably a year and possibly longer. For the long-term benefit of getting it right, this was felt to be a worthwhile investment. The suggestion on how to proceed was to bring together a working group of Operators with likeminded people who are motivated.

"Needs to be done very carefully and precisely."

"If it (the data) is not consistent and well defined, it is difficult to get accurate research."

"You cannot underestimate the amount of work that will take. It will need to be a team of data science specialists."

Some collaborative work has already been instigated on data projects by the Gambling Commission, the Remote Gambling Association and the Senet Group with varying degrees of progress. All the interviewees who had any participation with these projects described the definitions issue as a barrier to progress and indeed the Senet Group as part of their response to the Gambling Commission's consultation on its new National Strategy to Reduce Gambling Harms stated the following:

> "Senet supports the development of an industry data Repository, but its recent experience in leading a data collaboration project would suggest

[6] http://ris.leeds.ac.uk/ris/info/70/ethics

> that this is a complex challenge, where the comparability of data will necessitate clear research objectives."
>
> Senet Group 25th February 2019

Our advice on this issue is to ensure that there is a close collaboration between the organisation setting up the Repository and Operators providing the data from the outset, with the inclusion of engaged Researchers who will use the data. It is easier to set out clear goals and agree on definitions at the outset rather than to reverse engineer these. Most of the agreements that the CDRC sign with data providers are the product of detailed discussion between data provider and Researchers who will use their datasets.

Uncertainty over data requirements

One major barrier raised by the Operators was around uncertainty over the type and quantity of the data required. Their point was if it is top level aggregated data it will be too superficial for Researchers to undertake any meaningful work whereas on the other side as they (the Operators) have so much data then the provision of it could be overwhelming and a waste if only a small percentage of it is used. Their view on how to overcome this matter was for specific research questions to be set to which they could then supply the appropriate data.

This approach was advocated by several of the interviewees and illustrated by the following quotation.

> *"What sort of data would it be that you are after …..because without knowing the type of questions that are going to be posed of it, we could provide and recommend a set of data but it does not answer any of the questions a researcher may have in their minds."*

As with producing consistent definitions, this process needs to be undertaken collaboratively by Operators, the organisation running the Repository and Researchers.

Solutions for dealing with definitions and data barriers

The interviews with Operators and the Gambling Commission raised four different models which may help to deal with these dual issues around consistent of definitions and the provision of data.

1. A team with representatives from different Operators is convened to develop a set of definitions that can be used across the industry.

One of the Operators was enthusiastic about this approach having been involved in the building of a central database within another industry and the creation of a Master Data Management document within their own business. This Master Data Management schedule has 40 definitions relating to each customer and thereby illustrates the potential size of the definitions issue across the industry.

Two Operators with recent experience of collaborating with others saw this as needing a very long timescale to achieve. One of whom was unenthusiastic about this way forward and favoured a more tailored research orientation.

2. A set of Guidelines on definitions is produced by a central body (unspecified at this stage) and provided to all Operators with a view to their adapting their data to match these Guidelines.

This approach received a rather neutral response from many of the Operators with the idea being seen as a possible solution but with its own complications. Plus, with a natural reluctance to having anything imposed, it was received with limited enthusiasm.

3. Asking Operators to provide data to be used in the answering of specific research questions.

This would initiate the process of creating common definitions, give the Operators a specific task to complete rather than open ended discussions and provide them with the confidence to know the exact purpose for which their data is being supplied. Over time this would be extended and the practices and knowledge gained would be built into widening the data held within the Repository. There was wide ranging support for this option.

"Build it (the Repository) up by degrees and extend it upon well understood reasons"

"Starting small is good and it can grow gradually and so becomes a natural journey"

A view expressed by some of the Gambling Commission interviewees was that Operators should see organising their data into the required formats as a cost of business.

4. Extend the existing returns required by the Gambling Commission with the data required to be submitted to be increased over time. Furthermore, if it is a requirement that has to be undertaken on a regular basis (monthly/quarterly) then it is more likely to happen.

Whichever option is adopted, the Operators were keen that action should result from their provision of data i.e. research would be undertaken and they were motivated by seeing valuable outcomes rather than becoming involved with long discussions that absorb energy and resources with no foreseeable benefits to the Industry and its customers.

In our experience, often the best way to obtain data is by consultation between the Operator and Researchers who have specific questions they want to answer. This is most akin to option 3 presented above. The initial relationship can be built upon as research questions develop. There is no quick fix to this and the data which are supplied are often a compromise between the supplier and the research team.

Summary

Dealing with the barriers identified by Operators is key to the success of a national Repository. Key issues relate to:

1. Legal, data protection and ownership of data
2. The consistency of terms and datasets across different Operators
3. Identifying what data would actually be required

The first of these issues can be dealt with by drawing on existing models of data sharing. We have provided an overview of CDRC practices and emphasise that a strong legal agreement between the Operator and the Repository is the foundation of a sustainable relationship between data supplier, holder and user. The second issue was seen by many Operators as the key barrier and will require substantial investigation in order to ensure consistency in the definitions used in data which is deposited in a Repository. The final issue around the types of data needed does not have an immediate answer. On this issue, we hope that the associated *Patterns of Play work*, being undertaken by NatCen will shed some light on specific data requirements. We do suggest this work needs to be undertaken in consultation with both the Operators and Researchers to ensure that there is consistency between what can be produced and what can be used for research purposes.

One of the options presented by Operators for dealing with definition and data requirement issues was to build a Repository slowly, asking Operators to supply data for answering specific research questions. This is an approach which we would advocate. Dealing with these concerns require a high degree of trust between data supplier and data holder, and we expand on this in the next section.

6 Building trust with data partners and users

Operators interviewed were primarily interested in achieving results and to maintain their commitment to such a project then it will need to be seen to be moving forward in the short term in addition to offering longer term solutions. Users are keen that access to a Repository is straightforward, with no lengthy delays to the process or undue administrative burden. In this section we outline some of the ways the owners/administrators of a Repository can build trust with both Operators who are willing to supply data and with users of those data.

Engaging with data partners

Underpinning much of the success CDRC has had with engaging with data suppliers has been the Centre's strategy around data acquisition, otherwise known as its 'Ladder of Engagement', shown in Figure 1. In 2017 the Ladder of Engagement was endorsed by the Economic and Social Research Council (ESRC) as an approach which should be adopted across the Big Data Network.

The overarching objective of the CDRC's data acquisition strategy has been to engage with consumer facing organisations as providers of data for academic and other users (including government, third sector and the general public). Such providers are recognised as occupying the top tier in a hierarchy of collaborations. The lower tiers can be viewed as 'rungs on a ladder' through which the process of data acquisition is managed strategically from initial contact to full and open data sharing. Organisations at the lower tiers (partners, prospects and participants) also generate value in their own right, often helping to create networks, entering into joint projects, sponsoring PhD students or supplying data for internships.

Getting organisations to the top of the ladder of engagement has been challenging. Developing trust has been essential and has often involved being able to clearly and convincingly communicate to our partners the importance we place in safeguards and legal processes. Demonstrating how sharing can be of benefit to the organisation has also been critical and has often been achieved at lower rungs on the 'ladder' through joint ventures and/or the offer of training. Working with organisations on joint projects, offering insight on their data and practical solutions to business problems, as well as presenting opportunities to link different domain data, has proved particularly helpful in signing up partners. Where propositions have been clear and focused, partners have responded particularly well.

Tier 1 **Providers**	Characteristics: Data is made available to a wide constituency of end users. Access could be open so that the data is freely available to any third party; or restricted e.g. to protect commercial interests. Value: Multilateral data sharing for academic and external users.
Tier 2 **Partners**	Characteristics: Bilateral collaboration between one of the CDRC research groups (or contribution to MSc or intern programme) and an external organisation. Value: Leads to case studies and/or peer-reviewed publications.
Tier 3 **Prospects**	Characteristics: Meetings have taken place with organisations leading to a concrete proposal under consideration. Value: Generates insight regarding the provenance and content of partner datasets.
Tier 4 **Participants**	Characteristics: Interactions have taken place with those who may act as gatekeepers, or are strong advocates for collaboration, and with whom some degree of preliminary engagement has been established. Value: Facilitates business understanding and raises awareness.

Figure 1. The CDRC Ladder of Engagement

In the context of setting up a national Repository for gambling industry data, we would not envisage that all Operators would reach Tier 1 from the outset. Issues around attitudes to sharing data, the expertise, resources and infrastructure available to Operators and the volume of data they hold will all dictate the level of early engagement. It would be sensible to focus resources and energy on getting a few highly engaged Operators to the top of the ladder, but working with the rest of the industry to engage them at some level.

User access to the Repository, governance and user training

Access to data held within a Repository needs to balance the competing requirements of safeguarding potentially sensitive information vs making access relatively easy for those who have a legitimate reason for using the data.

Deciding who should be allowed access

Researchers and Operators alike recognised that the Repository would need to be managed in order to provide information and access to enquiries, although opinion on the level of access differed. Overall, opinion tended to coalesce around the position that the data within the Repository (at least initially) should be made available just to bone fide Researchers. This prompted discussions about how Researchers may be defined with the general conclusion being this must be an area that Universities and Data Base Holders are used to managing. User access is dealt with in the CDRC by implementing 'user journeys', depending on the type of data which are being sought.

User Journeys

In the CDRC, access to safeguarded and controlled data has been configured around two separate 'user journeys', both of which have been streamlined over time and which begin with the completion of an online, initial application. Initial proposals are processed by either the CDRC Senior Management Team (SMT) at the University of Leeds or UCL depending on which institution is the primary custodian of the requested dataset. If the initial application is approved, users are then supplied with a full proposal form and assigned a Data Scientist to act as an advocate for the project. Full proposal forms for controlled data are slightly more involved and need to be accompanied by a CV. Once received, both safeguarded and controlled data applications are assessed for remit fit initially by CDRC SMT. If approved, they are then passed to the data partner and a member of Research Approvals Group (RAG) for review. The data partner has a right of veto and if an application is rejected at this stage it will go no further in the approvals process.

Upon SMT and data partner approval, safeguarded data applications are then sent to the RAG Chair (see below) for auditing purposes only. Controlled data applications, once approved by SMT and data partner, are then sent to two academic peers (preferably with domain expertise) before being passed onto the RAG Chair for final approval. Processing times for safeguarded applications range from 2-4 weeks (depending on any revisions that may be requested), while controlled applications take in the region of 6-8 weeks to process from receipt of initial application through to data release.

Governance

Interviewees expected there must be precedent for the governance of access to data and this would be a question for the "governance panel" to decide upon once it had reviewed the regulations of other existing models/databases. Generally it was believed the process should demonstrate independence with a founding principle of its governance that no one should receive preferential treatment. It was hoped by all parties that when data is requested from the Repository the rules or "governance panel" would be designed to enable timely access rather than creating a long process. There was a view that governance procedures should not stifle innovation.

We received feedback on the types of Governance questions which should be put to those requesting data including:

- Who are the users wishing to access the data?
- What credentials do they have?
- Do they have clear research objectives and questions?
- Is the data they want to access relevant to their research?

It was the view that those requesting access need to be *"qualified people and they need to have a justification or a user case".* It was suggested that access applications could be reviewed on a "Levels Based Approach". This is akin to the service levels used in the CDRC (open, safeguarded, secure).

It was also assumed that Researchers accessing the data would be part of a team with their own data mining expertise and so the administrators of the Repository would not be expected to undertake extensive work on behalf of the Researchers.

The CDRC Research Approvals Group (RAG)

Many of the comments around governance are covered by the CDRC Research Approvals Group (RAG). It was a requirement of the ESRC that the use of CDRC safeguarded and controlled data be permitted independently of the Centre's Senior Management Teams. The RAG was established to satisfy this requirement with its membership comprising of a Chair, independent academics in the research field and a CDRC Data Scientist in an ex-officio capacity. Data partners involved in projects are also co-opted RAG members who review proposals in advance of other RAG members in order that permission for the use of data is determined at the earliest possible stage. The RAG's composition ensures that there is sufficient expertise in research design and analysis, as well as in policy impact. The RAG plays a key role in ensuring the quality and scientific merit of research being conducted through the Centre. It typically considers applications remotely which has enabled the Centre to return decisions on user applications in a timely manner. To receive RAG approval, project proposals must demonstrate:

- *Scientific advancement* – how the project has the potential for the project to advance scientific knowledge, understanding and/or methods using consumer data;
- *Public good*- how the project has the potential to provide insight and/or solutions that could benefit society;
- *Privacy and ethics* – the potential privacy impacts or risks, and wider ethical considerations relating to the project;

- *Project design and methods* – how the project will be conducted and who will be involved with a focus on demonstrating project feasibility;
- *Cost and resource issues* – what impact the project is likely to have on CDRC resources, including staff time and use of infrastructure, as well as any data acquisition costs. Resource requirements should be justified.

RAG approval will not be granted without evidence that the applicant has obtained ethical approval for the research through their institution, or else supplied justification as to why such approval is unnecessary. For non-academic projects, these requirements are assessed by the CDRC institution at which the data are to accessed and analysed.

The CDRC SMT(s) role in initially reviewing full proposals is primarily to determine the feasibility of any project in terms of our capacity and capability to deliver. Once an application has been approved by the RAG, CDRC senior management are then required to issue User Agreements to all those who will be accessing data.

Safe Researcher Training

Users accessing controlled and, in most cases, safeguarded data are required to have successfully completed some form of safe researcher training. The precise nature of that training is specified within the User Agreements and is largely determined by the data partner in question.

For access to safeguarded data, many data partners require users to complete the online MRC training course, *Good Research Practice: principles and guidelines*. Though the training was developed for those undertaking medical research, the principles taught around ethics and sound research practice apply effectively across all research domains. This training is free and can be completed in an hour or two. Other data partners are satisfied that users complete their institution's GDPR training, while others still place no requirements on training. Ultimately decisions on training are taken on a case by case basis by partners, often times upon the advice of the CDRC.

For access to controlled data, all data partners currently require users to have successfully completed safe researcher training provided by the UK Data Service, ONS, ADRN or HMRC. This training, *Safe User of Research data Environments (SURE),* is valid for five years and takes a full day to complete. CDRC has generally helped to facilitate access to such training and has relied predominantly on the UKDS and ONS for support in this area. UKDS charges students £30 (Others £60) for their one-day SURE training, whereas the ONS training is free. Going forward the CDRC is looking to deliver this training in-house to speed up and support uniformity in administrative processes.

Summary

Key to building a successful Repository is establishing trust with both data providers and data users.

- We suggest an approach which does not try to treat all suppliers of data the same, but rather engages Operators in a tiered way. This involves strong engagement with a smaller number of engaged partners who are prepared to share data and to engage other Operators in other activities. We term this approach the 'Ladder of Engagement'.
- Robust processes are needed to assess which users should be granted access to data housed in the Repository. Key to building trust with users is that this process should be robust, timely and equitable.
- Devolving decisions about access to a governance group (in our example a Research Approvals Group) who have clear guidelines makes this process robust and transparent.

7 Data acquisition and infrastructure

The infrastructure and processes required for hosting and managing data are not insignificant. Processes for obtaining, cleaning, storing and making data available need to be well defined and documented so that it is transparent to both suppliers of data and to users. Here we describe some of the logistic challenges and solutions for dealing with commercial data.

CDRC Infrastructure and Processes for on-boarding data at Leeds

The Integrated Research Campus (IRC) is the enabling technology platform that supports the CDRC Data Service at the University of Leeds. It is an advanced computational infrastructure that is highly secure and scalable to meet the needs of data-intensive research using personal and sensitive data. The IRC has attained accredited certification to the international standard for information security management ISO/IEC 27001:2013 and the NHS IG Data Security Protection Toolkit.

The IRC platform is an isolated "walled-garden" environment, with data entering and leaving via an "airlock" where disclosure control and information classification can take place in a controlled manner by the IRC Data Services Team. Data transfers are handled by a SFTP platform that uses end-to-end-security. Both safeguarded and controlled CDRC data assets are stored on the IRC, as well as other data supplied to the university, for example health data from NHS Digital, and crime data from the police.

Data Curation

Data provided to the CDRC often arrives in formats less than ideal for analysis using statistical software. For instance, a series of annual or monthly reports that are easily digested by a human but ill-suited for a computer. In some cases, data may be received as a longitudinal set of flat files with inconsistent content due to changes to source systems over the course of the data collection. In other cases, it may be the line of business source systems export data in a format that some Researchers may not be familiar enough with to handle effectively, such as *.xml or proprietary files. In all such cases, the Data Services Team are on hand to transform the data into a clear, non-proprietary and consistent format before making them available for research. Data are also quality assessed, for example inconsistency with previous downloads or missing fields can be identified early and queried with the data provider.

This process highlights that concerns identified earlier around data definitions are not constrained to the data which might be supplied by gambling industry Operators. Much of the data we receive is messy, and requires considerable resource to clean and make consistent. This needs to be factored in to the resourcing of any Repository: some of this work could be done at source but someone needs to be responsible for making sure data are useable by the research community.

The Data Services Team, working with data partners and Researchers, are also responsible for the creation of data profiles that are made available via the CDRC data portal. These data profiles contain a record of relevant metadata, for instance, a background to the scope, scale and methods of data collection, some descriptive statistics and, where possible, a research case study.

CDRC Secure Service at Leeds – Infrastructure and Processes

Users seeking access to the CDRC secure service are only able to access controlled data through the IRC secure environment where the risks of data leakage can be mitigated. Users access these data via safe rooms which are locked and can only be accessed by permitted users. Recording devices, such as phones, cameras, laptops and notebooks, are prohibited from the safe rooms.

Safe rooms are booked in advance with project granularity; users accessing data for different projects are not permitted to be in the same safe room simultaneously. Users can only access their Virtual Research Environment (VRE) from the safe room remotely through a thin client. VREs exist on the IRC and help to ensure the confidentiality, integrity and availability of data for each project. VREs enable remote access to project resources in a constrained and secure environment running with approved applications, rather than data being released.

The Data Services Team control all data flows to, from and within the IRC; there are technical constraints to prevent data flow via other routes. Before any data is released from the IRC Researchers are asked to describe the content, source, method of generation and disclosure mitigation measures that have been taken. The files themselves are then checked against these responses and against the terms laid out in all applicable data sharing agreements. Outputs are released only if the files are within the agreed acceptable threshold for disclosure. No source data leaves the secure environment.

UBDC and eDRIS
As stated previously, UBDC has been able to offer a secure service by using the facilities of an established and highly respected third-party provider, the electronic Data Research and Innovation Service (eDRIS). The benefits of turning to a trusted external service is that UBDC have not had to invest significant time or money in establishing infrastructure and in recruiting specialist personnel to oversee the handling of sensitive data. UBDC has also been able to pass on the considerable legal and ethical responsibility of data safeguarding to eDRIS. There are, however, some potential drawbacks to outsourcing a secure service, one being the additional layer of consultation, contractual obligation and administration that is involved. eDRIS, for example, need to be consulted at various stages of project development and approval, contracts need to entered into between eDRIS and researcher institutions and between eDRIS and users, and all of this no doubt results in a considerable

CDRC Safeguarded Service - Infrastructure and Processes

Users of the CDRC safeguarded service are granted access to the data via remote download from the IRC SFTP platform. Only users who have been invited by the Data Services Team to register for use with the system can access the platform, and only direct recipients of data packages can access the deliveries.

Principles of minimisation are applied whereby only data items required for the stated research purposes are released. Users can use the results of their analyses in publications, reports and presentations provided they abide by the terms and conditions of the data partner. There is no screening of outputs by CDRC staff.

CDRC – Safe Data Handling and Data Services Support

The Data Services Team act as trusted third parties to Researchers and data partners. Data transfers to and from the IRC are handled by the Data Services Team who check the data for compliance with data sharing agreements and other applicable contractual and regulatory obligations. Any non-conformities are handled by the Data Services Team before anything ever leaves the secure environment or can be accessed by the Researchers.

The Data Services Team also provide support to execute data transformations to ensure data is available to Researchers in a format most useful to them. This includes, but is not limited to, creating analysis files derived of raw data and performing aggregations to reduce data sensitivity and de-identification of personal data. The Data Services Team can also act as a liaison between Researchers and data partners to ensure research requirements are technically feasible and achievable.

De-identification at source (see Pseudonymisation) is a valuable process for enabling the release of information for secondary purposes such as research. Pseudonymous linkage operations are required where research relies on such de-identified information from more than one source being linked on the level of the individual.

Linkage between datasets brings numerous issues, even when data is pseudonymised at source. Risk of re-identification must be addressed, and this must be particularly stringent where information is sensitive. Linkage should neither involve the flow of more than the minimum dataset, nor enable information about individuals to flow between data sources. The Data Services Team provide assistance with the linkage of data, often pseudonymised at source, and sometimes by the team, to address such issues.

Data Sourcing Service

One process which has been supported by UBDC but not currently at scale by CDRC is researcher requests for data not currently held within their collection. This Data Sourcing Service is largely supported through UBDC's data acquisition budget, although many Researchers will often incur some costs in seeking new data, for example, in relation to licensing and/or data preparation.

To access this service, Researchers must make an initial online application. If internally approved, Researchers are then asked to submit two documents: a project proposal form and a data acquisition form. UBDC data scientists are assigned to a project in order to support the development of both documents.

UBDC only progresses new acquisitions that it believes are of strategic value to the broader research community. The DAC evaluates data sourcing applications based on additional criteria, including, data availability, data life expectancy, legal and ethical issues, level of technical support required to make the data accessible.

This sourcing service is worth mentioning as it might help to support a more bespoke approach to building relationships with data providers, in a way which is researcher led.

Summary

This section has provided some of the technical detail about how the CDRC on-boards and deals with data.

- The amount of staff time required to ensure that data are clean and consistent is considerable. Much of this is done by CDRC but potentially some of this work could be done in-house by the data provider.
- A secure infrastructure platform (and associated processes) is vital for the integrity of data and in maintaining trust between suppliers and users of data.
- A data service staff is also an important part of a Repository as they work with providers and users as a trusted third party, ensuring that data are robust and that outputs are appropriate.

8 Costs of establishing and running a Repository and service

This section sets out some of the costs of setting up and administering a Repository which adheres to the principals set out in earlier sections of the report.

Costs presented here are based on the business model of the Consumer Data Research Centre, funded by the ESRC. In terms of size of engagement, these costs are based on a successful model which has:

- data agreements in place with 26 organisations who are sharing data with Researchers at a national level;
- over 50 organisations engaged in various joint working activities;
- received 300 applications for data held in secure and safeguarded tiers;
- over 12,000 registered users (the bulk of whom download our open data (>225k data downloads).

These costs can only be used as a guide, as a Repository of gambling industry data will have very specific requirements. However, the number of partners who share sensitive information and the number of users accessing those data could be seen as indicative of a long-term goal for a Repository, given that there are over 1,000 licenced Operators in the remote gambling sector alone. We include some information of the rationale for each of these costs.

Table 1. Indicative costs for running a data service (per annum)

Ref.	Expenditure item	CDRC like service £	CDRC 'light' £
A	Investigators (senior management)	26,569	26,569
B	Business Development	90,720	70,728
C	Management	77,414	66,782
D	Software Engineering	188,385	102,044
E	Administration	94,521	50,222
F	Data Acquisition	90,000	60,000
G	Travel and Subsistence	13,500	9,000
H	Infrastructure platform	165,000	110,000
I	RAG Chair	5,000	5,000
J	Web development costs*	10,000	10,000
K	Training costs	7,500	6,000
L	Recruitment*	4,000	2,000
M	Indirect Costs	143,172	112,048
N	Estates	19,028	14,892
	Total	934,809	645,285

*these items of expenditure would not be required annually

In Table 1 we present the data service and infrastructure costs for mirroring the Leeds CDRC service. The second column provides a ‘pared back’ version of these costs which includes little in the way of engagement and outreach activities. These costs do not include any provision for research and cover only the administration of a data service. It should be noted that these costs might potentially be reduced if a Repository were aligned with an established Data Centre where infrastructure and processes are in place. More detail on items (A-N) in Table 1 are laid out below.

- **A.** **Investigators (Senior Management)** (totalling 2 days per week), are responsible for strategic oversight, operational leadership, management of staff and building research collaborations.
- **B.** The **Business Development Manager** engages with external partners, helping to ensure an ongoing flow of new data resources and project opportunities to keep the research active and relevant.
- **C.** The **Centre Manager** has operational responsibility across the project and is responsible for all planning and reporting tasks, coordinating resources, dealing with legal arrangements and delivery plans.
- **D.** The **Information Governance Manager** has oversight of the processes and procedures to maintain the integrity of data sets at all service levels. **The Research Software Engineer (RSE)** identifies resources necessary for the provision of data to projects and are responsible for the curation of new data sets including creation of metadata and data profiles. The **Senior Data Analyst** is responsible for the provision of data including coordinating access to computational resources and configuration of specialised resources such as Virtualised Research Environments or Cloud services. The **Data Scientist** supports the RSE and Senior Data Analyst in the curation, provision of data, and analytics for promotion and dissemination. The **Research Data Analyst** is responsible for the production of analytics to support dissemination (excluded under 'CDRC light').
- **E.** Administrative support (ADM) contributions range from supporting data acquisition and contracts through coordination of governance and user journeys to dissemination. **Two ADM posts** are complemented by a **Senior Administrator**.
- **F.** **Data acquisition** – We found that most external partners are not motivated by financial gain in seeking to share data, but that there are circumstances – in particular when creation of the data requires substantial effort – in which a modest budget is necessary to finalise an agreement.
- **G.** **Travel and Subsistence budget** is necessary for business development meetings.
- **H.** **Infrastructure Platform** costs are incurred in three main areas: the provision of Virtual Research Environments (VREs) for secure data projects; access to resources beyond the standard provisions for desktop networks (storage, memory, computation) in the Data Centre; and the introduction of computation in the cloud for both processing and provision of data.
- **I.** We pay the **Independent Chair of our Research Approvals Group** an honorarium of £5,000 per year.
- **J.** **Web development costs** are needed to create a robust infrastructure.
- **K.** Providing the right skills for infrastructure development and keeping those skills updated is crucially important so a **training budget** is included.
- **L.** **Recruitment** costs for advertising and promotion of new positions.
- **M.** **Indirect costs** are levied by the University for research activity and covers personnel services, library & computing services, central & local financial, administrative and technical management.
- **N.** **Estates costs** are levied by the University for things like heat, light, buildings and grounds

Optionally: **Researcher time** is funded by the CDRC. This cost would be relevant if there is motivation for the gambling industry data Repository to have research staff who can use the datasets and produce outputs. Each researcher would cost approximately **£44-50k p.a.**

Support for long term strategic funding

The view of most interviewees is that a National Repository has to be seen as a long-term strategic objective. If the energy, commitment and resources required to establish the Repository are to be invested then the opinion was expressed that it should be created for the long term and therefore its management and maintenance costs should be sustainable.

It was pointed out by the Researchers that to date a fundamental block to widening the research pool is the lack of funding. It was seen that, historically, there has only been sufficient funding to finance smaller short-term projects rather than longitudinal studies and until this is addressed then the creation of a Repository may only have a limited effect on the number of research projects that can be conducted.

> *"There is no point in creating this massive data Repository if there is no funding resource to be able to support it or create a community of people to make use of it"*

It was pointed out by the Commission that this situation is changing with the current National Strategy to Reduce Gambling Harms research programme containing big projects that intend to accelerate progress in a number of areas, including a longitudinal study. Nonetheless the Repository will need to be well funded in order to ensure its long term sustainability and some potential funding models are presented below.

Research council and match funding

Funding for a repository might come from one of the UK research councils (e.g. the Economic and Social Research Council). CDRC infrastructure and research activates are largely funded by ESRC but the centre also leverages funds from elsewhere, including commercial providers of data who wish to fund specific projects.

One option is to undertake match funding, where Gamble Aware, the Commission, the Operators or a combination of all three offer to match fund a research council investment.

Schemes which might be suitable include:

- The ESRC centres competition, funded at £2.5 - £10 million (100% of full economic costs (fEC) over five years.
- ESRC open call. Funding ranges from £350,000 to £1 million (fEC).
- ESRC Secondary Data Analysis Initiative (SDAI) funding (up to £200,000). This funding stream may fit the requirements of the Repository in the initial phase, of establishing the relationship between Researchers, Operators and the host of the Repository. It would require Operator data to be deposited within an existing ESRC investment, such as the CDRC. See section 10 for more detailed discussion.

Cost recovery models

As an ESRC funded centre, the CDRC data service has been free of charge to users. However, in future, and in the interest of sustainability, it is likely that researchers will be expected to cost service fees (e.g. for staff time, data acquisition, infrastructure, software) into their grants, or else be charged for data access to secure and safeguarded service tiers.

While UBDC's data services are, in general, free of charge to UK researchers and policy committees, additional costs are sometimes involved for data acquisition, cleaning, programming or linking, especially when sourcing new data.

A cost recovery model makes sense when demand for services is high. Users typically work with the provider (of the Repository) to work out how much staff, computing, storage and other resources are needed and apply for this through whatever funding stream they deem most appropriate, usually as a direct cost within their grant proposal to a research council.

Ongoing contribution from the Operators and other funding partners

Additional funding might come from the Operators via the voluntary levy on gambling profits or by using some regulatory settlement funds[7] which can form part of a settlement offer to the Gambling Commission. The Commission pointed out that there is work ongoing to increase the amount of funding that is provided by operators. These options should be considered in the context of a Repository forming part of the wider body of research in to minimising the harms of gambling.

Summary

We have presented here some indicative costs for establishing and running a Data Centre, based on the CDRC. We draw a similarity, at least in terms of absolute numbers, to the size of what a Repository of gambling industry data might eventually look like.

- It is likely that the inclusion of a Repository within or alongside an existing research centre would reduce costs, given that infrastructures would already be in place and a Repository could leverage economies of scale.
- To emphasise findings laid out earlier in the report, adequate provision of staff with required expertise is essential for the effective running of a data resource and their value should not be underestimated.
- There is wholescale support for an investment which establishes a Repository which is sustainable in the long run.
- In terms of funding, an investment will be needed for the setup of a Repository. Ongoing costs will need to be covered to ensure the repository is sustainable. Options presented include research council funding, cost-recovery models and contributions from Operators via voluntary contribution or by channelling some of the funds from regulatory settlements. In reality, some hybrid approach will likely be most appropriate.

[7] https://www.gamblingcommission.gov.uk/PDF/statement-of-principles-for-determining-financial-penalties.pdf

9 Starting, funding and sustaining a repository

In this section we make some specific recommendations for next steps in creating a repository. The steps outlined draw on the evidence collected for the report which strongly suggests that working with highly engaged stakeholders and utilising existing infrastructure where possible would facilitate and smooth the establishment of a Repository. It should be stressed that setting up a Repository is a complex undertaking so an iterative approach where all stakeholders involved (Operators, Researchers, the Repository host, etc.) can be flexible and think creatively about overcoming challenges is recommended.

The first subsection focuses on the short-term activities which need to be undertaken. We also discuss funding models and 'crowding in' activity which is one route to producing critical mass in research activity necessary for a sustainable repository.

A short-term plan for establishing a repository

This section sets out some simple steps over the short term for establishing a Repository. In this context we don't specify a time period for these short-term activities because this phase is focused on relationship building, and the timeline is not easy to quantify at the outset.

Step 1a. Work with a select group of highly engaged Operators and Researchers who want to answer a specific set of research questions. Identify the data required to undertake the research and enable collaboration between Researcher and data provider to establish how data sharing would work. This should be undertaken alongside the *Patterns of Play* research project which will have identified some of the suitable datasets available.

Step 1b. Engage with an existing Data Centre within a University to deposit relevant Operator data within their systems and draw upon existing infrastructure. The Data Centre will liaise with Operators and Researchers to identify issues referenced in this report around transfer, storage, anonymisation/pseudonymisation, licencing, IP, IG and access requirements. This phase will allow the Data Centre to iron out a series of complex issues in collaboration with Researchers and Operators. Such a relationship would require funding for infrastructure and staff support (see below).

Step 2. Continue to engage with Researchers, Operators and Data Centre to:

- Ask additional research questions of deposited data;
- Actively include further interested researchers, who can use deposited data;
- Make deposited data available for other research users to stress test data sharing infrastructure, for example by inviting applications for the datasets;
- Acquire further data from the Operators, based on discussions about further research objectives with the engaged stakeholders.

Step 3. Invite other engaged Operators to contribute similar datasets, again with a focus on answering some specific research questions. Repeat step 3 to increase the number of research questions being pursued, Researchers using the data and data available.

Step 4. Continue to expand activity, consider sustainability of repository in terms of funding and critical mass of research activity. At this stage the status of the Repository as an independent entity is important. While in the early stages (Steps 1-3) it would be sensible to draw heavily on existing Data Centre infrastructure while establishing processes, the Repository now needs to begin to operate self-sufficiently. In reality this means it has its own infrastructure (hardware, staff) and governance structure (established during the building phase). It should maintain strategic alignment with the Data Centre, drawing on best practice and working collaboratively to scale up research activity.

Attaining critical mass

In this section, we refer to strategies for attaining critical mass required for a successful and sustainable research centre. This is referred to as 'crowding in' because it includes activities which can leverage the Repository to undertake connected activities. These are strategies used by the CDRC.

Aligned research projects are an important strategy in attaining critical mass. In this report we outline the mechanisms which exist for researchers to engage with the CDRC and use its data assets. A successful Repository will need users to make applications to use the datasets it contains. Many of these applicants come with their own funding for their research time, but the cost to the Repository of supporting this research activity needs to be taken in to account.

Internships have proved extremely valuable in undertaking substantive data science research[8]. We employ an intern for 12 months, and the intern will undertake two six-month projects in collaboration with an academic and, usually, a data partner. Interns are highly skilled and have ambitions to become professional data scientists after their internship. Each six-month project costs £18,000 to provide for intern salary costs, intern training, travel and subsistence and admin costs. One option for a Repository would be to mirror this scheme, with a data partner (or academic researcher) contributing to the cost of the internship, which can utilise Operator data held within the repository.

PhDs provide the basis for long-term research within the university system. Typically three to four years, the long-term outputs can be substantial as a postdoctoral student can spend time on a single, focused topic. PhDs require funding, which can be contributed from commercial partners, from research council funding in the form of stipend and fee awards, or paid for by the individual undertaking the studentship. Tuition fees for home students are set by Research Councils UK (RCUK) and the 2018-19 rate is £4,327 per annum for full-time research degrees (for UK nationals). Additionally, a stipend award usually includes living costs, around £15,000 per annum.

[8] https://lida.leeds.ac.uk/study-training/data-science-internship-scheme/

Short-term funding

In the report we discuss longer term strategies for funding a repository. However, in the short-term the model which would enable a Repository to be established without reliance on external funding bodies would be to:

- Seek a contribution from the Operators; and/or
- Establish a recharge model, for researchers at the point of access

Both of these approaches have advantages and disadvantages. An Operator contribution might provide the most immediate source of funding, but might not be sustainable in the long term. This would also need careful consideration in order to retain independence for the researchers, as it is important that the data repository infrastructure is perceived as being independent and free from industry influence. The most obvious way to achieve this is to use the funds from the voluntary levy on profits, channelled via the National Strategy to Reduce Gambling Harms. A recharge model would require Researchers to have some funding (e.g. institutional, research council or consultancy) to commit and it might be difficult to obtain this commitment at the inception stage of a Repository.

Another option might be to collaborate with highly engaged Researchers, and the University housing the Repository to apply for Research Council funding to undertake some specific research, for example through the Secondary Data Analysis Initiative.

Secondary Data Analysis Initiative (SDAI)

The Economic and Social Research Council administer the SDAI[9]. Funding is provided for up to 18 months with an overall limit of £200,000 per grant and typically around 20 proposals a year are funded. SDAI will only consider projects which seek to exploit in innovative ways one or more of the ESRC data resources – of which the Consumer Data Research Centre is one. So as a solution for short term funding, an option might be to deposit some data within the CDRC (or other ESRC funded research centre) and leverage the SDAI to undertake some well-defined research activity. It is important to stress that the SDAI is a competitive funding mechanism and from submission to award can take circa 12 months. This is a solution which would benefit from a collaborative approach between Researcher and Operator.

Collaboration with a research council

Another option might be to seek to collaborate with a research council, in conjunction with the University hosting the Repository. A recent example of such a collaboration was the CDRC's ESRC Innovation Fund Initiative. Launched in 2017, this initiative invited calls for proposals from UK-based academics with the aim of establishing a network of projects that would utilise the Centre's core datasets and broaden its network of partnerships with academic, business, public and third sector organisations. The Innovation Fund ultimately supported ten projects across a broad range of domains, each valued at approximately

[9] https://www.cdrc.ac.uk/research/secondary-data-analysis-initiative/

£50,000, and was successful in both promoting the Centre's assets and in cementing longer-term partnerships with external organisations.

Summary

This section has outlined a series of initial steps for setting up a Repository. We suggest working with an existing Data Centre and engaging with an Operator and Researcher(s) who can help tackle many of the challenges which the Repository will need to deal with.

In the medium to longer term, the Repository will need to attain critical mass. We spell out the mechanisms for doing this, inviting researchers to apply for and use data, the provision of internships and PhDs. Funding might come from operators or at the point of access from Researchers. We also suggest that working with an established Research Centre there could be an opportunity to apply for Research Council funding, the most obvious currently available is the ESRC administered Secondary Data Analysis Initiative (SDAI). The SDAI would require Operator data to be deposited within an existing ESRC funded resource (such as the CDRC or UBDC) and the Repository could be built from there.

10 Conclusions and Recommendations

In this report we have provided a summary of evidence gathered for a scoping study setting out the feasibility of setting up a National Repository of Gambling Industry data.

There was wide ranging support for establishing a Repository from all of the groups interviewed: Operators, Researchers, The Gambling Commission and Gamble Aware. It was thought that such a Repository could help alleviate issues identified by Researchers including a lack of funding within the UK for gambling related research and the lack of access to data. Operators recognised the benefit of sharing their data with a national Repository to enable robust research in to a range of issues, including the harms of gambling. They also saw that outputs from research supported by the Repository offered the potential to provide results and benchmarks which could be used within their business. From the point of view of the Gambling Commission, the establishment of a Repository would drive new thinking and increase the volume of evidence-based research.

The answer to where a Repository should be housed and who should be responsible for maintaining it was not immediately obvious to stakeholder groups, although after discussion it was largely agreed that a university or group of universities would be a suitable option. This is because universities offer sufficient intellectual rigour when dealing with the data and posing research questions, and because they provide the most obvious link to the research community who would use the data.

Key issues in the setting up of a Repository were identified. All parties anticipated there would be difficulties in gathering data in a uniform manner as each Operator has their own definitions relating to the descriptions of customers and their actions. Operators also have different data structures and processing facilities. This inconsistency was seen as a key challenge for the creation of a national Repository. Uncertainty about data requirements for the Repository was identified by Operators as a key challenge and there was a call to ensure that any data supplied was used and useful, essentially warning against asking for large volumes of data which were subsequently found to be irrelevant. In terms of logistics, the challenge of setting up legal agreements, data protection and ownership of data were identified, but it was felt that these could be dealt with by borrowing from existing good practice, for example the processes used by university Data Centres.

None of these challenges are insurmountable, and we have presented in this report solutions, largely based on our experience running the CDRC, which engages with a wide range of commercial organisations and makes their data available to Researchers. The CDRC, and other examples of Data Centres cited in this report (e.g. UBDC, the UK Data Service) demonstrate that there is substantial appetite from both commercial organisations and Researchers to make use of data for research purposes. In the next section we present a set of recommendations to be taken forward in the establishment of a national Repository.

General Recommendations

In this section we make a set of recommendations for the establishment and housing of a Repository of gambling industry data based on the evidence detailed within this report.

Start with a smaller number of highly engaged stakeholders who can help build the Repository. Be prepared to engage with Operators at different levels rather than expect all to be able to share data immediately.

Of the various solutions discussed by Operators for setting up a Repository, the one we recommend is to start with a number of highly engaged stakeholders and build up the repository from there. This would allow issues around consistency and data requirements to be ironed out collaboratively and provide some data relatively quickly which could be used for research purposes. We advocate an approach similar to the CDRC 'Ladder of Engagement' (see Section 6), whereby committed Operators are on the top rung, making data available through the Repository, while other operators are engaged on lower rungs, through activities such as answering discrete research questions in collaboration with Researchers. This way, within the framework of a Repository, there is an opportunity to engage with a range of Operators and users, with potential to build relationships and increase engagement over time.

As far as possible, work collaboratively to agree on definitions and data requirements.

There was large scale agreement form all stakeholder groups that having a collaborative approach, between Operators, Researchers, Gamble Aware, the Gambling Commission and other parties to dealing with obstacles was a good way to proceed. An approach where requirements are imposed on Operators without consultation was unsurprisingly received with little enthusiasm from the Operators. We advocate a collaborative approach as far as possible, where stakeholders have a say in how the Repository is established, as obtaining and housing data requires trust between all parties. However, there is likely to be some nuance to the process and ultimately the requirements for the Repository to define and store data and processes which are useful and appropriate for the research which needs to be undertaken should take precedent.

Establish strong legal agreements between Operators and the host of the Repository for a sustainable relationship.

Many of the legal and governance issues of sharing data can be dealt with early on in a relationship with Operators by establishing clear and robust legal agreements, which set out how data are to be stored and how they can be used. We outline in Section 5 the three legal agreements which underpin the CDRC data service. Whenever possible, we would recommend that standard licence agreements templates be used as this allows for ease and consistency of implementation and clarity and consistency around user journeys and obligations.

Establish the Repository within a university, or at least ensure that a university is closely aligned with the work.

When pressed, stakeholders engaged as part of this project highlighted that universities would offer the scientific rigour to be able to legitimately operate a Repository. Universities provide a key link with Researchers who will use the data. We also suggest that universities provide independence from the gambling industry which is important in providing legitimacy in an area of research seen by many as difficult because of perceived ethical issues.

Align with an existing Data Centre to reduce costs, draw on expertise and borrow established governance structures.

In order for a Repository to be sustainable it needs to be adequately funded and have highly engaged data providers and users. The set up costs are substantial but ongoing costs need to be taken into account, particularly with regards to providing adequate staffing. Some of the set up and ongoing costs may be reduced by aligning a Repository with an existing Data Centre where infrastructures are already in place. This would also have the added benefit of providing established governance structures and processes which already work. Our recommendation would be to look at strategic alignment with an existing centre. Alternatively, a Repository should borrow heavily from processes which have worked in such centres.

Ensure that a Repository has adequate staff who can represent the Repository as a trusted third party.

Having an adequate number of staff who can effectively run a Repository was identified by interviewees and highlighted in this report when discussing the necessity to ensure data are consistent and useable. There is also a wider requirement which is to have experienced staff in place who can represent the repository as a trusted third party, building relationships with Operators, Researchers and other stakeholders. The value of this role should not be underestimated because building and maintaining relationships goes beyond just the infrastructure of data storage.

Player data needs to be linked across different Operators in order to provide in-depth research on the harms of gambling.

Establishing the exact data requirements for the Repository needs further work, but one aspect which was emphasised strongly by Researchers was the requirement to link data across different Operators. This is key to undertaking substantial research, especially where players have multiple accounts with different providers. This linkage needs to be discussed in detail in the requirements gathering exercise and methods for linking these data needs careful consideration. We discuss potential methods for linkage and issues around disclosure in Section 5 of this report.

Specific short term recommendations

Here we make some specific recommendations for immediate next steps.

Work with a group of highly engaged Operators, Researchers and an existing Research Centre to define research questions, establish processes and iron out problems.

Working with a select group of highly engaged Operators and Researchers who want to answer a specific set of research questions will quickly highlight some of the requirements (from a research point of view) and possibilities/limitations (from a data provider point of view). Engaging early with an established Data Centre will enable the appropriate infrastructure to be built from the ground up while leveraging existing expertise and infrastructure during the building phase while keeping costs down.

Scale up the Repository to attain critical mass and become self-sufficient.

Once the repository, its processes and principles have been established by the engaged stakeholders, scale up to include additional Operators and Researchers, but continue to work iteratively: leveraging additional data and posing new research questions. Crowding in other activities can help attain critical mass for the Repository: e.g. align with other (funded research projects), introduce internships to undertake short term research projects and PhD studentships for longer term projects. Ensure the Repository is self-sufficient in terms of funding and infrastructure.

Obtain immediate funding from Operators and seek to leverage Research Council funding.

One source of immediate funding could come from the Operators via the voluntary levy on gambling profits or by using some regulatory settlement funds. Additionally, research council funding could be sought to cover some of the costs (e.g. the Secondary Data Analysis Initiative, administered by the Economic and Social Research Council).