

Novice to Know-How Module Text

Course 7: Providing Access to Preserved Digital Content

Module 6: Managing Sensitive Data

The development of this course was funded by The National Archives (UK) as part of the "Plugged In, Powered Up" digital capacity building strategy.

1. Introduction.

This module addresses some of the main ethical and legal issues relating to sensitive data, explaining how they can limit or restrict user access to digital content at your organization.

It provides an overview of the ethical and legal issues for dealing with confidential, personal or other sensitive content. It is written from a UK perspective with references to relevant UK legislation. Other relevant legislation from outside the UK will also be mentioned to highlight commonalities and offer general guidance for managing sensitive data.

Please note that the information and guidance offered in the module do not constitute legal advice. The module aims to provide general summaries and guidance for legal issues. The authors recommend you seek legal counsel for your specific circumstances and guidance for individual requirements.

2. What do We Mean by Sensitive Data?

Sensitive data can mean many things depending on the context. Personally and ethically, we can think of it as any information that someone does not wish to share or that might negatively impact them or others.

General definitions of terms you will come across when working with forms of sensitive data are listed below. They are based on UK legislation but can generally apply to all organizations.

- **Personal Data.**
 - Personal data generally means any information relating to an identified or identifiable natural person or, in other words, anything that is clearly about a particular person. In certain circumstances, this could be anything from a person's name to their appearance.
- **Sensitive Personal Data.**
 - Sensitive personal data, or 'special category data' in UK legislation, generally refers to categories of personal data that should be treated with extra

security. It can include (but may not be limited to) data about a person's race or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, physical or mental health condition, genetics, biometrics, health, sex life or sexual orientation. It may also include data on a commission or alleged commission of any offence, proceedings for any offence committed or alleged to have been committed, and disposal of such proceedings or the sentence of any court in such proceedings.

- **Confidential Data.**

- Confidential data generally means any information given in confidence or agreed to be kept confidential among the parties involved, and is not in the public domain. Some confidential data will also be personal data and/or sensitive personal data, but it may also include data outside these categories that have been agreed upon between parties.

Sensitive personal data is not limited to just the creators or donors of materials in our collections. It is important to remember that sensitive data, especially in research data, can include other individuals or groups who are represented or identifiable in the content.

3. Legal Definitions of Terms.

We have provided general definitions of terms, but there are legal definitions to follow in the context of providing access to preserved digital content in your organization.

This will again depend on your legal context, but those provided below are example definitions of 'personal data' from the EU General Data Protection Regulation (GDPR) 2016, and from the UK General Data Protection Regulation (UK GDPR) 2018 and UK Data Protection Act 2018 (that share the same meaning of the term).

- **EU General Data Protection Regulation (GDPR).**

- Personal data means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, and online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person

- **UK General Data Protection Regulation (UK GDPR) and UK Data Protection Act 2018.**

- Personal data only includes information relating to natural persons who: can be identified or who are identifiable, directly from the information in question; or who can be indirectly identified from that information in combination with other information. It may also include special categories of personal data or criminal conviction and offences data. These are considered to be more sensitive, and you may only process them in more limited circumstances.

4. Knowledge check (interactive slide).

Knowledge check question: True or False, Confidential data can include other kinds of data that do not fall into the 'personal data' or 'sensitive personal data' categories.

The correct answer is "True". Confidential data can include other kinds of data in addition to the 'personal data' and 'sensitive personal data' categories.

5. Key issues when managing sensitive data in digital content: Ethical Issues.

Opening up digital collections can support organizational transparency and openness, particularly for public bodies. However, not all information can be ethically or safely shared. Privacy and confidentiality are ethical considerations essential to building trust between your organization, donors, users, and other stakeholders.

You should discuss these ethical issues with donors and record creators as early as possible, ideally before or during transfer and ingest. This will inform them of how the organization has reviewed the ethical considerations of sharing sensitive data and assure them you will not share anything that falls under sensitive personal information without first obtaining their permission or consent.

This also helps establish methods of informed permission or consent. Discussing the likelihood of digital content containing sensitive information will give donors and record creators time to think about any data they wish to limit access to.

Make sure that you record the information gathered and agreed upon during these discussions in your digital asset register and other documentation.

6. Key issues when managing sensitive data in digital content: Legal Issues.

It is also important to identify, understand, and acquire all the legal permissions needed for managing, using, or granting access to sensitive data in collections. Otherwise, your organization may be at risk of breaking the law and receiving fines or other penalties.

Understanding which legislation applies can be challenging, even when the ownership of a collection as a whole is well understood, because there can be sensitive data about other individuals or groups in the content of materials within the collection.

The legal issues for sensitive data in digital content are arguably more complex than for analogue materials. Most legislation was established with the analogue in mind, and good practices for the preservation of digital content not always recognized or allowed by existing provisions.

As we will get into shortly, the nature of digital content also creates challenges of timing and scale for identifying sensitive data before opening access. Therefore, the management of sensitive data in content and collections often requires a range of more complex strategies and risk-based assessments.

7. Knowledge check (interactive slide)

Knowledge check question: Discussing the ethical issues surrounding sensitive data helps support (choose all that apply):

- Openness
- Transparency
- Trust
- Publicity

The correct answers are "Openness", "Transparency", and "Trust".

8. Some Examples of Relevant Legislation.

There will be different legislation for different kinds of organizations in different contexts, and legislation is frequently amended, so it is best to visit government websites to make sure you are accessing the latest versions.

Some examples of relevant legislation for the UK, Europe, and the USA are listed below. A list of these and legislation from a number of other countries is provided in the additional resources document for this course.

- **In the United Kingdom** there is the
 - Data Protection Act 2018,
 - UK General Data Protection Regulation 2018,
 - Freedom of Information Act 2000.
- **In Europe** there is the
 - General Data Protection Regulations 2016.
- **In the United States**
 - There is no one general federal legislation impacting data protection in the USA. Instead there are several federal data protection laws that are sector-specific or focus on particular types of data.
 - For example, the Freedom of Information Act (FOIA) 5 U.S.C. § 552.

9. Data Protection legislation.

Data Protection legislation protects the privacy and integrity of data held on individuals by businesses and organizations to ensure that individuals have access to their data and can correct it, if necessary.

In the UK, the Data Protection Act is enforced by the Information Commissioners Office (ICO), overseeing the Freedom of Information Act and the regulation of interception of communications under the Regulation of Investigatory Powers Act 2000 (RIPA). The Data Protection Act covers data held electronically and in hard copy, regardless of where data is held. It also covers data held on or off site.

10. Freedom of Information legislation.

Freedom of Information legislation provides public access to information held by public authorities, usually in two ways: public authorities are obligated to publish certain

information about their activities, and members of the public are entitled to request information from public authorities.

The UK Freedom of Information Act 2000 covers any recorded information held by a public authority in England, Wales and Northern Ireland, and UK-wide public authorities based in Scotland. Information held by Scottish public authorities is covered by Scotland's Freedom of Information Act 2002.

The legislation relates to members of the public submitting a data protection subject access request to sensitive or confidential information that a public authority may hold about them, so it may only be relevant if your preserved digital content includes this kind of information.

11. Time and scale challenges with digital content.

Providing access to digital content facilitates sharing and opportunities for learning and reuse but presents challenges for identifying sensitive information. In some cases, the consequences of accidentally releasing sensitive information in digital content can be very serious, potentially compromising human safety.

Whether it pertains to classified government information or human subjects represented in a research study, sensitive personal data must be identified. However, traditional manual review methods typically used for sensitive information in analogue materials do not scale well to the volume of data in digital content enabled by technologies. Additionally, examining the content of collections, files, codes, and other digital objects to ascertain whether they contain sensitive or potentially sensitive material is time-consuming.

12. Tools for identifying sensitive data.

Balancing the two objectives of sharing and protecting has created a demand for reliable approaches for automating the identification of personal and sensitive information within the volumes of data received by an archive or repository.

Digital content does offer new opportunities for automatically detecting the presence of potentially sensitive information. For example, there are tools and software like Bulk Extractor, Forensic Toolkit and EnCase Forensic, which can help identify sensitive information when undertaking processes. But for most organizations, there will still be processes in place to check the accuracy of automated tools before making content accessible.

13. Knowledge check (interactive slide).

Knowledge check question: True or False, it is good to check the accuracy of automated tools before making content accessible.

The correct answer is "True".

14. Managing and controlling access levels.

Before any access is granted, permissions or restrictions for sensitive data should be identified and documented.

The information you have captured in your digital asset register and resource discovery metadata will help you determine the appropriate level of access or no access at all. You can then apply appropriate access restrictions and processes to protect sensitive data.

Keep in mind that 'sensitive' is not a term designating a level of access but rather a term characterizing the content that should inform access decisions. You can use existing access models to assign and apply levels of access, for example those provided in the Levels of Born-Digital Access, which are:

- **Open:** Open materials are available for research with no known restrictions.
- **Conditional:** Collections that include both open material and material with restrictions, which may include materials that are deemed "sensitive" or under copyright.
- **Closed:** Closed materials are not made available to researchers. Materials that are closed may eventually be made available after an embargo period. Collections or materials may be closed if they contain information protected by applicable law or sensitive information; or if a donor has requested an embargo period.

As will be discussed in the next module, these access levels can also be used for managing copyright and other intellectual property rights.

15. Mediation and Authentication.

It is helpful to be aware of other options for the delivery and access of digital content. There are different options for mediating content, including authentication, to facilitate access.

For example, you could make 'open' materials accessible to users on the web via a digital collections website where they do not need to submit a request or login (i.e. unmediated and unauthenticated).

For conditional access to sensitive data, you would want to have different degrees of mediation or authentication. For example, sharing files via a cloud storage service or using encryption for off-site access to files through a secure website (HTTPS) and/or a Virtual Private Network (VPN).

When you are first getting started, there is no need to make these decisions regarding technical mediation and authentication straight away. Still, it is worth keeping them in mind for how you wish access systems or services to develop in the future.

16. What is Redaction?

Similarly, is it helpful to be aware of redaction as a method to limit access to sensitive data in digital content. Redaction refers to the process of analyzing digital content, identifying confidential or sensitive information, and removing or replacing it.

Common redaction techniques include anonymization and pseudonymization to remove personally identifiable information, and cleaning of authorship information. Depending on the resource, this could range from removing or blocking out names in documents to more complex removal of information in larger datasets.

17. Carrying-out Redaction.

When starting out, if you decide to use redaction, it is important to remember to:

- Carry out redaction on a copy of the original, never on the original itself.
- Assess whether data is sufficiently anonymized enough. One common test is the 'motivated intruder' test: you or another staff member think about someone who might be motivated to identify a person from (and gather personal data about) the information in the data to assess anonymization, determine risks surrounding the data, and decide if any additional measures should be taken to de-identify data or add security measures to protect data from unauthorized users.
- Be aware that personal or sensitive data can be collected that is not noticeable at first. For example Microsoft Office documents are stored in encoded formats that can contain information like change histories, audit trails, or embedded metadata that are not displayed and presence may therefore not be apparent.

While it is not critical to undertake redaction when first getting started, a helpful resource to explore redaction software is the DigiPres Commons and there is also The National Archives (UK) Redaction Toolkit (links are provided in Additional Resources).

18. Knowledge check (interactive slide).

Knowledge check question: The three types of access levels provided in the Levels of Born-Digital Access are:

- free, conditional, closed.
- open, conditional, closed.
- open, temporary, closed.
- open, conditional, prohibited.

The correct answer is "open, conditional, closed".

19. Four Actions for Managing Sensitive Data

There are two areas in the Levels of Born-Digital Access (LBDA) that address sensitive data issues and actions—Description and Security. Managing sensitive data also falls under the Legal Basis area in DPC RAM and arises at various points in the Access workflow.

The next slides bring together the guidance provided by these resources for four basic actions to take at your organization:

1. Analyze your content and assess compliance,
2. Document critical information,
3. Apply access or use restrictions accordingly,
4. Provide information and user guidance on conditions of access and reuse.

20. Action One: Analyze your content and assess compliance.

Start by reviewing relevant policy, guidelines and supporting documentation, and associated metadata from transfer and ingest to identify sensitive or potentially sensitive data. If it is not already, privacy and sensitive data should be part of a larger policy or a standalone policy, and there should be written policies relating to related rights. Institutional processing guidelines will need to be sufficiently flexible to respect the various kinds of categories or restrictions that may exist for content.

Next, undertake a risk assessment based on relevant legislation to assess compliance in regard to sensitive or potentially sensitive information. Determine whether access restrictions or other ethical considerations have been discussed with donors and whether further information or discussion is needed.

From the basic assessment of compliance based on this analysis, you can establish or improve access controls and security measures.

Remember that risk assessment and decision-making should be an iterative process, so if you have already accomplished this action, you still may benefit from undertaking it again—especially if legislation has undergone any recent changes.

21. Action Two: Document critical information.

The second action is to document important information about sensitive data, legislation and compliance you have collected during the previous action.

Note that some materials may require an agreement before access, and others may have access restricted or closed at the owner's request. The use of templates for transfer agreements and/or metadata will help streamline processes and documentation for the above, and make sure that all the critical information is recorded in your digital asset register and metadata for other discovery or access systems at the organization.

22. Action Three: Apply access or use restrictions accordingly.

The information you have captured will determine how, or if, the content should be accessed by users. It can help streamline the identification of digital content with potentially sensitive data for review and the application of restrictions, making access to the content open, conditional, or closed accordingly. There may also be digital content identified that can be accessed so long as restricted elements are redacted before access is provided to users.

Before enabling or restricting access, make sure that you have addressed

- What restrictions have been placed by a donor or related parties?
- Which digital content is subject to data protection restrictions?
- What access might still be possible given the above restraints? How can you adapt your access levels or models to facilitate this?

To reach Level One of Security in the LBDA, you should provide access to open, authentic, virus-free content on a dedicated on-site public access computer with security measures implemented based on the restrictions. For any content made accessible online, there must be a mechanism for individuals or organizations to request it be taken down due to a breach in law or ethical concern.

23. Action Four: Provide information and user guidance on conditions of access.

The previous module explained how user guidance is important to help users overcome the barriers they may face in accessing and reusing content of interest. This applies to sensitive data as well. You must be clear about the reason or justification for restrictions on the access, use, or reuse of content. Providing the correct information about restrictions or permissions for sensitive data in a way that is easy for users to find while searching or requesting digital content can help ensure that they fully understand their ethical and legal responsibilities.

As part of reaching Level One of Description in the LBDA, you should provide required descriptive elements for a collection-level record and at least one descriptive note about the processed digital materials. This should include a collection level explanatory note on 'Conditions Governing Use' that informs users of restrictions relating to sensitive data. You might also want to include administrative metadata elements pertaining to the status of sensitive data.

Terms and conditions should also be given to users when requesting material and when access is provided. This both informs and guides users on what is permitted or restricted by law, and their ethical and legal responsibilities.

24. Module Summary.

This module provided a broad overview of the ethical and legal issues surrounding sensitive data and access to preserved digital content.

In summary, it is essential to identify any sensitive or potentially sensitive data in digital content before resources are made discoverable and accessible to users. This includes:

- Discussing ethical and legal considerations with donors as early as possible to establish and maintain trust, and obtain permissions or conditions for access in a way that respects privacy and confidentiality.
- Identifying relevant legislation and assessing compliance to address any legal responsibilities of the organization for restricting or enabling access to personal data, and implement processes for managing access with roles and risk owners at the organization.
- Capturing and sharing critical information in resource discovery metadata or documentation so the information is easily findable and readable to users, enabling them to engage with the content in a legal and ethical manner.

While this module offers general guidance, your organization should consult with a legal advisor to address your specific circumstances and guidance for individual requirements.