Old Bailey Online: XML files

Author: Sharon Howard

Date: March 2017

[**Overview**](#_j6do3o2raa5r) **2**

[**Licensing and Citation**](#_o7t6ipn1yrn) **2**

[**The files**](#_saw0rl4czkyh) **3**

[**Identifiers**](#_locdhkiupbo7) **4**

[OBP trial IDs](#_x83b6h61imos) 4

[OBP Defendant IDs](#_ongsc1ylcdtm) 4

[OA person IDs](#_mkplvwhuhu7e) 4

[Other IDs](#_eoxrowsf8ods) 5

[Constructing Old Bailey Online URLs from IDs](#_2b2gl5wh73g) 5

[**XML tagging**](#_4kup7rhdeyzi) **5**

[OBP: Outline XML structure](#_iiycyxh29z25) 5

[Tagging of names, offences, verdicts, sentences](#_5al0q1odmrgu) 6

[Linking defendants, offences and outcomes (etc) with <join> tags](#_bmlc7h5boxco) 6

[**Offence, verdict and sentence categories**](#_pe0kgx1ggvkq) **7**

[Offences](#_ncc4ekw6xx92) 7

[Verdicts](#_6awufnugalud) 9

[Sentences](#_g748winti3tq) 10

## 

## 

## Overview

Further information about the project, historical background and a searchable version of the data is available at <http://www.oldbaileyonline.org>

The technical methods used to create the data are described at <https://www.oldbaileyonline.org/static/Project.jsp#methods>

The XML files are marked up using TEI, and therefore in general users should refer to the TEI documentation (<http://www.tei-c.org/>). This document is intended only to explain and describe specialised aspects of the markup and data structures not covered by the generic TEI schema.

The XML markup of the Old Bailey Proceedings data is highly complex, reflecting its dual purpose and the inherent complexity of the source. The first purpose was to produce full-text transcripts of trial narratives that could be many thousands of words in length. The second was to enable quantitative analysis of long-term patterns in prosecution and punishment. The complicating factors are, firstly, that a single trial may involve **multiple defendants, multiple offences, multiple verdicts and multiple sentences**, and secondly that these elements of information do not necessarily appear close together in the text, or in exactly the same position in every trial, or in exactly the same format. (Verdicts and sentences *normally* appear at the end of a trial, but some verdicts and sentences might be recorded earlier in the trial, and for part of the 18th century sentences were recorded in a separate section after the trial reports.) Ensuring that defendants were correctly associated with offences and outcomes was often not a straightforward task.

XML can represent this kind of complex data model very effectively without needing to separate full text and structured information, helping to ensure the integrity of the data associations. However, the markup is extremely verbose, with complex nested structures, and so the files are not easy to work with. This document aims to help make the data more accessible, but there are no quick fixes: you will need to invest some time and effort in order to understand the data and use it effectively.

The Old Bailey Online project cannot provide unpaid technical support for users of this data. However, we do wish to encourage research and re-use of the data as much as possible. Specific queries (as well as ideas for funding proposals!) may be directed to [oldbailey@sheffield.ac.uk](mailto:oldbailey@sheffield.ac.uk) but be aware that we cannot guarantee that we will be able to help.

## Licensing and Citation

All Old Bailey Online data is released under a Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC 4.0).

<https://creativecommons.org/licenses/by-nc/4.0/>

Note that our interpretation of 'commercial' normally excludes academic research outputs, even when published with commercial publishers. Please contact the email above if in any doubt.

Please refer to the citation guidelines in the website for general principles:

<https://www.oldbaileyonline.org/static/Legal-info.jsp#citationguide>

The most appropriate citation form for research outputs using data derived from this dataset is likely to depend on the exact nature of the new data, but should include the project name **Old Bailey Online** and the main website URL **http://www.oldbaileyonline.org**, the **version number** of the data, and individual IDs where appropriate. If the new data is also made publicly available we suggest the inclusion of a statement along the following lines:

This work builds on [Old Bailey Online](https://www.oldbaileyonline.org/) that is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License. Permissions beyond the scope of this license may be available at <https://www.oldbaileyonline.org/static/Legal-info.jsp>.

## The files

This document describes two datasets, distributed in a zipped archive (OBO\_XML\_7-2.zip)

1. Old Bailey Proceedings [**OBP**] part 1 (1674-1834) and part 2 (1834-1913) (2163 files)
2. Ordinary's Accounts [**OA**] (475 files)

Accompanying documentation:

* OB\_xml\_notes.docx (this document)
* listOBP.txt
* listOA.txt
* licence.txt

The current version of the website and its data is **7.2** (March 2015).

Each OBP file represents **one session** of criminal trials held at the Old Bailey. Each OA file represents **one pamphlet account** of criminals sentenced to death at the Old Bailey and executed at Tyburn, published by the Ordinary of Newgate.

Parts 1 and 2 of the OBP files were created and processed in two separate phases of the project: the first phase (2000-05) digitised Proceedings from April 1674 to October 1834; the second phase (2006-08) digitised the Proceedings from November 1834 to April 1913. This division reflected a key historical change to the court itself, when it was renamed the Central Criminal Court, its geographical jurisdiction was expanded, and the number of sessions per year was increased from 8 to 12.

There are no substantial structural differences between parts 1 and 2, but there are some differences which can affect XML parsing: for example, defendant names in part 2 are often enclosed in <hi rend="smallCaps"> tags that are not present in part 1.

The OA data is less complex than OBP, but it does contain markup for linking person names to other kinds of information (occupations and places).

See the following for further information on publishing history:

* <https://www.oldbaileyonline.org/static/Publishinghistory.jsp> (OBP)
* <https://www.oldbaileyonline.org/static/Ordinarys-accounts.jsp> (OA)

## 

## Identifiers

It is crucial to understand the IDs used in the XML files for various entities, which usually appear as attributes within the XML tags. The starting point for all IDs (and file names) is the **date of the session** of the Old Bailey (the opening day on the title page) or the **publication date** of an Ordinary's Account, in the format **YYYYMMDD**. Thus, for example, **18000219** identifies the Old Bailey session that began on 19 February 1800.

NB: Almost all of the identifiers within files were automatically generated, either in the initial phases of automated processing and markup or in post-processing. They are often sequential numbers (and may even appear to bear some relationship to numbers that appear in the original text). But if an automated markup process failed to correctly identify an entity (eg trial, defendant, offence), then IDs would need to be inserted manually. Therefore, apart from the session ID, it can never be assumed that any element of an ID has any particular meaning within the text.

However, the overall construction of IDs follows certain principles, of which the most important are outlined here.

### OBP trial IDs

These consist of the session ID and a unique identifier for the session in the form **tYYYYMMDD-{uid}**. Normally the unique identifier is numerical; but occasionally it will be a number with a letter added.

There are non-trial sections of text in many files which are similarly identified but with a different prefix letter, in particular **f** for front matter; **a** for advertisements; **s** or **o** for supplementary material / summaries.

### OBP Defendant IDs

These come in two forms, both incorporating the session ID, trial ID, and an identifier for the defendant.

**OBP Part 1**: {trial reference}-{defendant unique identifier}

Starts with a trial reference and and adds a text string, normally 'defend', and a number (which is unique within the file), eg: **t18000219-7-defend97**.

Occasionally the text string may be 'person' rather than 'defend', where the name was not initially identified correctly as a defendant.

**OBP Part 2** : {defendant identifier}-{trial number}-{session ID}

The second format uses a defendant number which is unique within the trial, followed by the trial number and the 8 digit session ID, all prefixed with 'def', eg: **def2-996-18400302**.

### OA person IDs

These are based on the pamphlet ID, paragraph number and a number for the name instance, eg **OA17000906n3-1**.

### Other IDs

In addition, victim and other tagged person names, offences, verdicts and sentences, dates and some other tagged entities, have unique IDs based on similar principles. These are used for linking tagged information, including associating defendants with offences and outcomes.

### Constructing Old Bailey Online URLs from IDs

The IDs are also used to construct URLs, which you may find useful for linking (although please note that we don't recommend using these URLs directly for citations in publications, as they could change in the future, whereas IDs are permanent).

**Session**

https://www.oldbaileyonline.org/browse.jsp?name=**{session ID}**

**Trial**

https://www.oldbaileyonline.org/browse.jsp?div=**{trial ID}**

**OA**

https://www.oldbaileyonline.org/browse.jsp?name=**{OA ID}**

## XML tagging

See the section on Markup in the project Technical Methods for a list of the types of information that have been tagged: <https://www.oldbaileyonline.org/static/Project.jsp#methods>

### OBP: Outline XML structure

<TEI.2>

<text>

<body>

<div0 type="sessionsPaper">

<div1 type="frontMatter">[title pages and (part 1 only) lists of jurors; (normally?) only 1 instance per file]</div1>

<div1 type="trialAccount">[one per trial]</div1>

<div1 type="supplementaryMaterial">[part 1 only: information about outcomes, normally 1 per file; not present in all files]</div1>

<div1 type="punishmentSummary">[part 1 only: information about sentences, normally 1 per file; not present in all files]</div1>

<div1 type="advertisements">[part 1 only, normally 1 instance per file; not present in all files]</div1>

</div0>

</body>

</text>

</TEI.2>

The structure for OA files is similar but simpler, with the <div0> @type "ordinarysAccount".

### Tagging of names, offences, verdicts, sentences

Person names (in OBP and OA) are marked up with the TEI **<persName>** tag, with **<interp>** tags used to provide further attributes such as given name, surname, gender and age. There are a number of different name types: defendants, victims, witnesses, judges and jurors. Both the name type and the name ID are indicated in attributes within the <persName> tags, eg:

<persName **id="def1-235-18530131" type="defendantName"**>

Offences, verdicts and sentences (in OBP) are marked up using TEI **<rs>** and **<interp>** tags. A simplified example of offence tagging (verdict and sentence follow similar lines):

<rs id="t18341205-201-offence-1" type="offenceDescription">

<interp inst="t18341205-201-offence-1" **type="offenceCategory" value="theft"**/>

<interp inst="t18341205-201-offence-1" **type="offenceSubcategory" value=**"**burglary**"/> for burglariously breaking and entering the dwelling-house of Samuel Raley 7th of December</rs>

(In reality, the victim name and offence date would also have further tagging.)

A full list of the categories and subcategories is provided at the end of this document.

### Linking defendants, offences and outcomes (etc) with <join> tags

There is *at least* one TEI **<join>** tag per OBP trial linking together the IDs for the correct offence(s), verdict(s) and sentence(s) for each defendant. (It will be present even when there is only one of each in a trial.) Example:

<join result="criminalCharge" id="t18380514-1139-charge-1" targOrder="Y"

**targets="def1-1139-18380514 t18380514-1139-offence-1 t18380514-1139-verdict-1"**/>

Further <join> tags are used to link other information relating to a particular offence instance, such as the offence date and the victim name.

The <join> tag is similarly used in OA to link people to occupations and places.

## 

## 

## Offence, verdict and sentence categories

OBP has 9 offence categories, 4 verdict categories and 6 sentence categories, most of which have further subcategories. These are listed in full below. You should, however, refer to the Old Bailey Online website for detailed explanations of the *meanings* of these category labels:

<https://www.oldbaileyonline.org/static/Crime.jsp>

### Offences

|  |  |
| --- | --- |
| **offenceCategory** | **offenceSubcategory** |
| breakingPeace | assault |
| breakingPeace | barratry |
| breakingPeace | libel |
| breakingPeace | other |
| breakingPeace | riot |
| breakingPeace | threateningBehaviour |
| breakingPeace | vagabond |
| breakingPeace | wounding |
| damage | arson |
| damage | other |
| deception | bankrupcy |
| deception | forgery |
| deception | fraud |
| deception | other |
| deception | perjury |
| kill | infanticide |
| kill | manslaughter |
| kill | murder |
| kill | other |
| kill | pettyTreason |
| miscellaneous | concealingABirth |
| miscellaneous | conspiracy |
| miscellaneous | habitualCriminal |
| miscellaneous | illegalAbortion |
| miscellaneous | kidnapping |
| miscellaneous | other |
| miscellaneous | pervertingJustice |
| miscellaneous | piracy |
| miscellaneous | returnFromTransportation |
| royalOffences | coiningOffences |
| royalOffences | other |
| royalOffences | religiousOffences |
| royalOffences | seditiousLibel |
| royalOffences | seditiousWords |
| royalOffences | seducingAllegiance |
| royalOffences | taxOffences |
| royalOffences | treason |
| sexual | assaultWithIntent |
| sexual | assaultWithSodomiticalIntent |
| sexual | bigamy |
| sexual | indecentAssault |
| sexual | keepingABrothel |
| sexual | other |
| sexual | rape |
| sexual | sodomy |
| theft | animalTheft |
| theft | burglary |
| theft | embezzlement |
| theft | extortion |
| theft | gameLawOffence |
| theft | grandLarceny |
| theft | housebreaking |
| theft | mail |
| theft | other |
| theft | pettyLarceny |
| theft | pocketpicking |
| theft | receiving |
| theft | shoplifting |
| theft | simpleLarceny |
| theft | stealingFromMaster |
| theft | theftFromPlace |
| violentTheft | highwayRobbery |
| violentTheft | other |
| violentTheft | robbery |

### 

### 

### Verdicts

|  |  |
| --- | --- |
| **verdictCategory** | **verdictSubcategory** |
| guilty |  |
| guilty | chanceMedley |
| guilty | insane |
| guilty | lesserOffence |
| guilty | manslaughter |
| guilty | no\_subcategory |
| guilty | pleadedGuilty |
| guilty | pleadedPartGuilty |
| guilty | theftunder100s |
| guilty | theftunder1s |
| guilty | theftunder40s |
| guilty | theftunder5s |
| guilty | withRecommendation |
| miscVerdict |  |
| miscVerdict | noAgreement |
| miscVerdict | no\_subcategory |
| miscVerdict | postponed |
| miscVerdict | unfitToPlead |
| notGuilty |  |
| notGuilty | accidentalDeath |
| notGuilty | directed |
| notGuilty | fault |
| notGuilty | noEvidence |
| notGuilty | nonComposMentis |
| notGuilty | noProsecutor |
| notGuilty | selfDefence |
| specialVerdict |  |

### 

### 

### Sentences

|  |  |
| --- | --- |
| **punishmentCategory** | **punishmentSubcategory** |
| corporal | no\_subcategory |
| corporal | pillory |
| corporal | privateWhipping |
| corporal | publicWhipping |
| corporal | whipping |
| death |  |
| death | burning |
| death | deathAndDissection |
| death | drawnAndQuartered |
| death | executed |
| death | hangingInChains |
| death | no\_subcategory |
| death | respited |
| death | respitedForPregnancy |
| imprison |  |
| imprison | hardLabour |
| imprison | houseOfCorrection |
| imprison | insanity |
| imprison | newgate |
| imprison | no\_subcategory |
| imprison | otherInstitution |
| imprison | penalServitude |
| imprison | preventiveDetention |
| miscPunish | branding |
| miscPunish | fine |
| miscPunish | forfeiture |
| miscPunish | militaryNavalDuty |
| miscPunish | no\_subcategory |
| miscPunish | sureties |
| noPunish |  |
| noPunish | no\_subcategory |
| noPunish | pardon |
| noPunish | sentenceRespited |
| transport |  |