**The Writable ‘Occupations’ Browser**

**Overview**

The tool described combines the features of:

1. a BROWSER:
   1. to explore the network of nodes (tokens/types) that describe an ‘occupation’ from an historical corpus, with each node representing an element in the description that is or will be reconciled to a series of taxonomies/vocabularies (based on established historical taxonomies, including a development of Booth-Armitage ‘occupations’, as used in simplified form for the Keele crowdsourcing project, material [Corfield], place [bespoke, derived from Locating London Lives Gazeteer].
   2. to define time periods as a filter for all views of the data set, using an interactively ‘bracketed’ timeline (dashboard component), and to provide a responsive histogram summary of the distribution of relevant data across that timeline
   3. To configure widgets representing different taxonomical types/subtypes, within a horizontally draggable menu/table of type-widgets, and to organise these horizontally to configure a sequence of node-types
   4. A Viewer within which to represent the Ego Network of a selected Node-Type, arranged (in its first iteration/version) as a centred entity-type node, and a ‘radial dendrogram’ visualisation of its subtypes, up to six layers deep, with the layers of subsidiary nodes arranged in concentric circles
   5. To manipulate the visualisation - browsing and selected nodes - to define a set of related entities/types as a filter/query for the visualisation of other dimensions of the corpus data
2. an EDITOR:
   1. To **accept** or **reject** raw token-sequences as feasible ‘occupations’ candidates
   2. to **split** (joiner /and/ carpenter), **compound** (Dog and Duck) or **group** (‘oil’ and ‘oyl’) tokens and token-chains
   3. to **assign** these tokens, token-groups, token-compounds to different types/sub-types within the taxonomies
   4. to **confirm**, **revise** or **delete** existing type-assignments (both generated manually and automatically)
   5. To further **specify**the type-assignment of tokens, token-groups or token-compounds that have been provisionally assigned (e.g. to re-assign from a higher to a lower, more specific level in the hierarchy).

The two modes are not distinct but occur within the same interface and involve interaction *between* and user interaction *with* the same (visualisation ‘dashboard’) components. However, each mode emphasises certain distinct features and functionality.

Development of the tool will prioritise the ‘editor’ functions, in its initial phases, but within the context of the overall design spec.

The test-case for the development of the browser uses the descriptions of occupations from the Old Bailey corpus. However, the tool is intended for broader application in the modelling, editing and browsing of the ‘active graph’, which mediates an ontological model of relational meaning in historical source texts and taxonomical records of semantic elements in those texts.

**COMPONENTS OF THE UI**

* **Taxonomy-Category/Node-Type Configuration Widgets**

Semiotic of the Taxonomy-Category Widgets

* + Widgets represent different taxonomical types/subtypes
  + Each widget presents via two fields: for type/subtype of the give taxonomy
  + Where taxonomies have several levels (more than two), additional widgets may be added to ‘fine-tune’ these levels (see below for details/restrictions)
  + Widgets are colour coded by their taxonomy type: e.g. title, place, object
  + Two states of colour intensity indicate whether or not a Category/type is included in a node-sequence currently selected in the ‘browser’

Configuring Widget Fields

* + Fields are configured by text entry, auto complete, menu/submenu; (by ref to taxonomy)

Direct Manipulation of Widget Sequence

* + Each widget is part of a panel of taxonomy/type-widgets that are individually, horizontally draggable
  + User drags to organise the widgets horizontally to select the sequence of node-types (ie. node bands) in the ‘browser’, along the arms of the radial dendrogram
  + \*Note: Widgets may not be dragged to the left of widgets with the same colour/type, as this would place lower taxonomical levels in an illogical ascendency over higher)
  + The chosen node-sequence corresponds to the ordering or nodes-types in the radial Vis, which reconfigures in response to any chance in the top panel

Killing/Birthing Widgets

* + Widgets can be ‘birthed’/‘killed’ (>Max7) then given a type (if none of same type exists)
  + Where a widget is ‘birthed’ it first appears on the extreme right of the panel, i.e. at the lowest level. It may then be assigned a Taxonomy-Category.
  + New widgets must be configured for types within their taxonomy. If a widget for that category already exists, the new widget may only be configured with lower level types in the hierarchy.

Widget interaction with Draggable ‘Ghost’ Nodes

* + On selection and dragging, nodes within the Ego Network Viewer create a temporary ‘ghost’ of themselves that is moved, while the main node remains in position.
  + These node ‘ghosts’ may be dragged onto a widget, in which case the node/token is assigned the lowest-level type displayed in a field
* **Ego Network Viewer**

Graphical Representations of Ego Graph

* + The Ego Network viewer has two visual forms: ‘Radial Dendrogram’ and ‘Monadic’
  + For current purposes, only the ‘radial dendrogram’ is relevant (Monadic left for later phase)

Ego-centred Radial Dendrogram

* + The ‘radial dendrogram’ is centred on a Ego Node that may represent any entity in any taxonomy, as chosen by the user
  + Around the Ego Node, sets of related entity-types are arranged in concentric bands, ordered from the centre outwards to match the sequence of taxonomy-types in the widget panel
  + For each node in a band (including the Ego Node), a node will appear in the next-band-out for any entity-type in the relevant taxonomy with which the inner node share a graph edge - eg. for title/place, any place that occurs in a phrase with that occupation title - and will be arranged a similar radial position (i.e. adjacent to it)
  + In certain instances, multiple nodes in a band will represent the same entity-type (i.e. be copies of one another), with which multiple entity-types in the next-band-in share edges

Radial Ordering of Branches

* + Nodes are distributed clockwise around each concentric band, with the default organisation being alphabetical by the first letter of the first token in any node
  + (Additional organisations of the clockwise ordering will be available subsequently)
  + *An exception to the ‘banded’ display is made for stop-words in the original token-sequence that have not been assigned a taxonomy-type of their own (e.g. ‘splitting’ or ‘within compound’), which are displayed as micro-glyphs just inside the band for the node/token with which they are associated (eg. ‘In the’ ‘fish market’)*

Semiotic Variation in Glyphs

* + A number of different glyphs (see below) are used to represent different states/contents of nodes [according to the ‘contextual zoom’, in later phase]
  + These glyphs vary by shape, coloration of border, and inner-out glyph shading, as described in detail below
  + Where nodes are displayed as two concentric glyphs of the same kind, the inner represent relevant entries with ‘confirmed’ taxonomical status, the outer those with ‘provisional’ taxonomical status
  + Nodes are displayed in different colours, corresponding to the Taxonomy-category to which a node is assigned.
  + The size of a node-glyph is scaled to indicate the relative frequency of its co-occurrence (in the tokenised phrases) with other nodes displayed in the current browser: e.g. the relative frequency of co-occurrence of the nodes in a band (only) with the Ego Node, or with the nodes in a selected sequence between it and the Ego Node.

Direct Manipulation of Nodes

* + As described above, the appropriate selection of a node creates a ‘ghost’ version of the node, which the user may drag within the browser (click-hold-drag), without displacing the original node from which it was ghosted.
  + Dragged ‘ghost’ nodes will only interact with a limited number of other elements in the display, which highlight when the dragged ghost node is over them to indicate their potential responsiveness (on click-hold-drag-release).
  + One element that may be responsive to a ‘ghost’ node dragged over it is: another node, in a different branch of the radial dendrogram, of the same taxonomy-category => the nodes are ‘grouped’
  + Another element that is responsive to ‘ghost’ node drag-over is a Taxonomy Widget => a node is assigned whichever taxonomy type is the lowest level displayed in the widget’s text field
  + On completion of a drag-drop to define a new relationship, a dialogue will invite the user to further configure the newly defined relationship

Direct Selection of Nodes and Node Clusters

* + A togglable option allows the user to select All/None nodes related to the Ego (in current Browser), as a default from which to fine-tune the selection.
  + Individual nodes may then be selected and deselected by simple mouse click, which toggles their ‘selected’ state (**and** that of any nodes on edges directly between them and the Ego Node) on and off
  + Clicking-dragging when (initially) not over a node creates a selection marquee which will only select/highlight nodes situated on the firstconcentric band that it touches (nodes on other bands are unaffected) [Note: click-drag + shift: click-drag allows aggregated marquee selections]
  + *By default, every selection of node/s is for their ‘inclusion’ in any resultant query. A togglable option allows the user to select nodes instead for ‘exclusion’ from any query (for later development)*

Graphical De/Emphasis of Un/Selected Nodes and Edges

* + The edges between related nodes in the radial dendrogram are, by default, represented as thin mid-grey lines
  + When certain nodes are selected to create highlighted node-sequences, their edges become emphasised (dark-grey/black), while those of non-selected nodes become de-emphasised (lighter grey)
  + When edges are de-emphasised, the colour density of nodes between those edges is also reduced.

Replacement Egos and re-centring the Radial Dendrogram

* + On relevant selection (mouse/hover-halo/highlight-click), any node will re-centre the radial dendrogram on itself as the new Ego
  + In effect, this establishes a fresh query, but any node-entities present in both queries (including, inevitably, the new and old ‘ego’ nodes) animate from their previous to their new positions

Interaction between Widget Panel and Nodes/Node Bands in Browser

* + At any point, any node in the primary selected strand can be adjusted using the widgets in the top bar.
  + Reordering the taxonomy-category widgets in the top panel will move node-bands within the existing taxonomical branch (up or down, to select less or more specificity), in which case the ‘subsidiary’ nodes (further out on the ‘spoke’ from its original or final position will adjust accordingly)
  + Birthing/killing widgets in the widget panel will add or remove bands in the browser and cause the radial dendrogram to reconfigure according (animating nodes/edge into their new positions and relationships)
  + Change to a term in different taxonomical branch (in which case the node and all subsidiary nodes will adjust accordingly)

*Cognitive Zoom*

* *The principle that the narrower the frame of the information displayed, the deeper, more complex and more detailed can be the related and contextual data represented by the visualisation*
* *An example of this in relation to the Browser might be: that when five node-bands were showing (close to the maximum ‘out’ zoom), there was no distinction in the glyph shape, which only showed their distinctive shapes when four node-bands were visible, and only showed inner/outer (Confirmed/provisional) glyphs when at the three node-band zoom.*
* *\*Note: for the first phases of development, the ‘cognitive zoom’ function will not be implemented: the design will be refined in response to testing.*

Saved States, a Query Terms in subsequent ‘Stages’

Each change in the configuration of the Browser is saved as a ‘state’, with users above to move backwards and forwards through these states (and, at a later stage, to browse them within a directory and diagrammatic representation).

Individual states current set(s) of highlighted nodes within the browser - individual, partial or full sequences - may be ‘bookmarked’ as pre-set filters for queries in visualisations of other dimensions of the dataset (e.g. scatterplot distribution of occurrence)

These states can be named and reviewed, with descriptive annotations and colours added, which will appear in other visualisations. On selection within the browser, they will return you immediately to the visualised ego network state.

The record of individual states also preserves a record of the statistics for the relationships between those elements selected within the state (e.g. relative proportions of different entity-types).

•**‘Candidate’ Phrases**

* Most text/phrases uploaded to the Browser in tokenised form are also likely to have been (automatically or manually) pre-assigned to likely taxonomical-types
* Where this has not occurred and the tokens are ‘raw’, the only edges for each token-node are those between the tokens in the phrase, with those edges between sequential tokens having particular weight.
* In any representation of such phrases in the radial dendrogram, individual nodes and contiguous nodes without assigned types will be represented as unique instances of an ‘unassigned’ type
* *Phrases containing unassigned token-nodes will always appear in a form that preserves the original order of the nodes in the phrase*

**• Branch-End Indicator Widgets**

* The individual branches of the radial dendrogram carry a specificity of meaning: they offer the most precise descriptions of an occupation, which the user may wish to view or define as a query
* Consequently, it is necessary that certain features are available that can be targeted at individual branches individually
* These features/tools sit within a composite graphical widget that is displayed in a band just outside the current outermost band, positioned adjacent to the last visible node in that branch
* This ‘band’ of widgets is concealed by default, with only a trace of the band itself displayed. The widgets themselves are only revealed on mouse-over, when their persistent visibility can be toggled on-off.

Summarise contents of Next-Band-Out

* The nodes and edges that comprise any displayed ego ‘branch’ in the radial dendrogram may not represent the entire phrase of which those nodes/edges are elements.
* (Note: this is especially likely to be the case in later phases of development, once the cognitive zoom is implemented)
* One function of the widget is therefore to graphically summarise what further content the branch may contain, beyond that currently displayed - and to allow the user to zoom ‘out’ to reveal it
* The mini widget for this is a fan shaped glyph, segmented into the number of branched nodes in the next-band-out, which when selected will recongfigure the Browser to reveal that next band

Indicate ‘Split’ Phrases Associated with Current Phrase Branch

* Where an original ‘raw’ sequence of tokens contained a conjunction which was identified as a ‘splitting’ node (i.e. it indicated the combination of two, or more, occupations), the phrase elements on either side of it were annotated as distinct phrases, related by the individual to whom they pertained
* The widget indicates (with a glyph comprising a purple circle containing a ‘+’) that the branch in question references a phrase (or phrases) that that are associated with one or more ‘split’ phrases)
* Click-selection of the ‘+‘ glyph causes the ego-branches for the associated phrase/s to highlight and become selected, with their edges coloured purple
* **Text Viewer/Annotator**
  + The text viewer shows a scrollable list of all those ‘occupations’ texts that include the nodes (token/token groups) and node-sequences currently selected in the Ego Network browser.
  + (The default ‘featured’ text - that displayed first - contains the most common tokens for the fullest current selected set of nodes, and the fewest other tokens.)
  + All token/token groups are highlighted/underlined in a colour corresponding to that of the node-type in the ego network (neutral - e.g. grey - for nodes yet to be assigned a ‘type’).
  + Clicking on a highlight selects it for manipulation (i.e. drag start-end of frame to change the words contained by it, ‘delete’ to remove the highlight)
  + Two check boxes at the right of the view (GreenTick/RedCross) allow the user to immediately annotate ‘candidates’ phrases as possible or to dismiss them as incorrectly proposed.
* **Timeline Histogram**
  + The timeline covers the period included in the corpus (in the case of OB c 1680-1840)
  + The brackets of a time-frame can be moved by the user to filter data by period
  + >> \*For current purposes, the timeline will be replaced by a text entry date range (e.g. 1740-1780)\*
  + The timeline histogram is the one ‘dashboard’ component that persists across visualisation ‘stages’ (within the overall Clioscope tool that will be implemented)
  + It represents the most simplified view of the timeline-as-scatterplot: stacked charts of occurrence over time, cross-referenced to/filtered by whatever elements are currently selected in Ego Network Viewer
  + Other organisations of the distributed timeline data will be presented in future design specifications (one example would be Ben Jackson’s ‘length of trial’ scatterplot)
  + Although a stacked chart, it is useful to conceive of the resultant histogram as composed of individual glypgs, which will animate/migrate into different organisations, as relevant

**The Data Structure (and Glyphs for Node Types)**

Underlying the tool is an ‘active graph’.

This graph represents the tokenised text of the corpus, organised and annotated by reference to:

Multiple taxonomies (2-4 layers deep) of entity types (relevant to ‘occupations’), top levels:

Role Title - **red**

Status - **orange**

Qualifier - **yellow**

Place - **green**

Object - **blue**

Person Name - **black**

Other (customisable: e.g. Conjunction/preposition) - **White/Purple** (by context)

Unassigned - **grey**

Vocabularies

That identify tokens/token groups (e.g. variant spellings) with a Type/SubType

An Ontology

Of types and subtypes and their (implicit, semantic) relations

[Note: the colour palette should ultimately be user customisable, with custom palettes available that are suitable for types of colour blindness]

Tokens may:

remain single - **circle** glyph (surveyor)

or be

grouped - **square** glyph. (e.g. oil and oyl)

compounded - **hexagon** glyph (Dog and Duck)

Tokens are persistently recorded as a linear graph of their sequence/token chain within the original text phrase in the corpus, as well as with reference to any groups, compounds and node-types to which they are assigned (which ‘inherit’ an annotation of their place within that token chain).

NOTE:

All glyphs default to **circle** when:

* The Ego Network is displaying (as foregrounded) more than 15 nodes (in total)

AND

* The Ego Network radial dendrogram has more than 3 concentric circles of nodes

UNLESS

• A node sequence is actively selected, in which case only its nodes are displayed in distinctive glyph form

Unassigned Tokens

Tokens or token groups that have not (yet) been assigned a Taxonimical Type (manually or automatically) have a particular ‘type’ status: ‘Unassigned’

‘Unassigned’ status determines their graphical representation as node-glyphs **(grey)**

Nodes with ‘unassigned’ status have limited functionality: eg. they cannot be selected as ‘ego nodes’ (on ‘hover’ no ‘halo/highlighting)

Provisional/Confirmed ‘Type’ Assignment

Tokens(and groups/compounds) assigned to types may be given *confirmed* or *provisional* tags.

*Provisional* tags - used, for example, for tokens that are grouped and given a ‘group name’ but without being assigned to a pre-existing taxonomical type - are used to surface these tokens and prompt users to refine their sorting to a type.

The ‘provisional’ tag is manifest variously within the tool, including: ’provisional’ nodes glyphs have no bounding line (in variant of main node colour)

Where a node-type glyph is scaled to indicate relative quantity/distribution, the quantities are graphically separated into core (confirmed) and outer ( provisional) elements of the glyph

Connecting/Splitting Tokens

*Where does the ‘And’ Go?*

*You can introduce a ‘split’ anywhere - ‘ands’ and punctation are represented as a prompt*

*— and ‘and’ between two identifiable different occupation titles will introduce a ‘split’ by default, though this can be removed (unless the combination of words is whitelisted: e.g. pub name)*

**Summary of Kinds of Interaction**

With A Node

* Hover - halo highlight - click => select as new Ego Node
* Click (no halo highlight) => select as node-in-chosen-sequence
* Click-hold-drag => create ‘ghost’ draggable node that may interact with other graphical events

• Dragover ghost node => highlight to ‘group’ + release mousdown +> ‘group’ with

With Other Glyphs

* Hover over extreme outer band marker => reveal micro-widgets related to spokes
* Click ‘fan’ widget element on extreme outer band => Expand to/reveal next band

• Click ‘+’ circle button => Select Split/Conjoined Sequence

With Widgets/Widget Panel

* Button => Birth/Kill
* Drag sequence of nodes => Order (with restrictions: cannot drag to left of higher level/same category)

**Possible Pre-Processing**

Tokenisation and stemming/splitting - by rules (e.g. split on an ‘and’ or punctuation that has two distinct role/title subtypes on either side of it)

Grammar matching to vocabularies, to generate *provisional* type-assignment at high-level of taxonomy only

Vocabularies similarly generate *provisional* token-compounds (e.g. for pub names)

All remaining, unassigned tokens are grouped with identical/near identical items as a ‘token group’ with ‘provisional' status and name of most frequently occurring token