ChatGPT

Representing Personal Data and Personality Traits: Existing JSON/Semantic Standards

Standards for Structuring Personal Content

ActivityStreams 2.0 (W3C): ActivityStreams 2.0 is a JSON-based format for representing social activities and content. It defines common object types like **Note** or **Article** for textual content (e.g. a journal entry or email) along with actors, timestamps, etc [1]. For example, a simple action "Alice posted a note" can be serialized with an **actor**, **object** (the Note content), and other metadata [1]. ActivityStreams is JSON-LD (Linked Data) friendly and highly extensible [2], making it feasible to add custom fields for psychological annotations. Using ActivityStreams, one could model diary entries, chat messages, or meeting notes as `Note` objects with properties like `content` (the text) and use the built-in `tag` array to attach tags for traits or topics. This standard is widely used in federated social apps (ActivityPub), so it's well-suited to **unstructured personal content**. Minimal extensions would be needed – for instance, defining a JSON-LD context for any new trait tags (ensuring they have a URI) [3]. Overall, ActivityStreams provides a flexible base for content with the ability to incorporate personality-related tags without breaking interoperability.

HL7 FHIR (Fast Healthcare Interoperability Resources): FHIR is a healthcare data standard, but it offers generic resources that can represent personal logs or survey data. The **Observation** resource in FHIR is essentially a name/value pair with metadata, meant for any measured or asserted information [4]. In practice, Observations cover everything from vital signs to subjective assessments, and even social or behavioral history [5]. For example, one could treat a journal entry as an Observation with a code like "Personal narrative" and the text as the value. FHIR Observations can also be multi-component, so one could include sub-components for, say, **derived trait scores** (e.g. an extraversion score extracted from that entry). Similarly, FHIR has a **QuestionnaireResponse** resource to capture survey answers (useful if the user took personality inventories like IPIP questionnaires) [6]. FHIR provides a rich structure (with standardized data types, coding of concepts, timestamps, author, etc.) and even a **Goal** resource for personal goals. The upside is strong interoperability (especially if integrating with health/psychology apps), but using FHIR outside clinical contexts may be heavyweight. It may require defining custom codes or extensions for niche concepts (e.g. a new code for a "meeting transcript" Observation, or extensions to mark a memory's significance). In a personality-cloning pipeline, FHIR could represent the inputs (diary texts, survey results) in a uniform way – e.g. treat each piece of personal content as an Observation with tags – but expect to do some profiling (customizing FHIR) to fully accommodate traits and adaptation metadata.

OpenHumans Metadata Approach: OpenHumans is a platform for aggregating personal data streams. Rather than a strict schema, it uses a **file + metadata** model: data files (e.g. a JSON dump of messages or a diary text) are stored with a **description and tags** provided by the uploader [7]. The tags are free-form keywords categorizing the content (e.g. `"mood journal", "survey response"`), and the description is human-readable. This approach is very flexible – virtually any personal data can be included – but it lacks a detailed structure or ontology. For a personality pipeline, OpenHumans would allow collecting all the raw data, but you'd need to impose your own schema on top of it to consistently extract traits. Essentially, OpenHumans gives you a **container** (with basic metadata like creation time, source, tags [8]) and you

could embed JSON inside files. It doesn't natively support semantic tagging of Big Five traits or narrative features; you would have to use tags or an external mapping. The advantage is compatibility with many personal data types out of the box, but the trade-off is minimal semantic detail in the default schema. This could serve as a **data integration layer** while another schema (like ActivityStreams or FHIR) is used internally for finer structure.

Other Relevant Standards: There are a few additional standards worth noting. The W3C **Web Annotation Data Model** (JSON-LD based) could be useful to tag pieces of text with semantic annotations. For example, one could attach an Annotation to a journal entry highlighting a sentence and tagging it with a trait like *"expresses high Openness"*. This model is compatible with ActivityStreams (they both use JSON-LD) and can use IRIs for tags (e.g. a reference to a term in an ontology). Another is **schema.org** – it provides types like `CreativeWork` (and subtypes `BlogPosting`, `EmailMessage`, etc.) to mark up content. Schema.org has properties such as `keywords` or `about`, which can be used to tag an item with topics or defined terms [9]. For instance, a diary entry marked up as an `Article` in schema.org could include a `keywords` field listing "BigFive:Openness, memory, family event" to describe its attributes. Schema.org itself doesn't define a vocabulary for personality traits, but it can carry those tags (the values could be plain text or URLs to a definition of the trait). In sum, schema.org is more lightweight than full ActivityStreams and can piggyback on existing web vocabularies, but it may require extension (e.g. a custom `additionalType` or use of `DefinedTerm` for trait categories) to formally represent traits.

Schemas and Ontologies for Psychological Traits & Metadata

When it comes to representing **personality trait data or psychological metadata**, there isn't a universally adopted JSON schema – but there are some building blocks we can leverage:

- **Personality Trait Ontologies:** Academic efforts have begun mapping personality concepts to ontologies. For example, Alamsyah et al. (2021) proposed an *"ontology-based model"* for Big Five traits extracted from social media [10]. Such an ontology would define classes for traits (Extraversion, Conscientiousness, etc.) and possibly link them to evidence (like words or behaviors). There's also work in the privacy community (OWL ontologies like the Data Privacy Vocabulary) that classifies *"personality traits"* as a subtype of behavioral data [11], underscoring that trait data can be formally described. While these ontologies are not yet mainstream standards, they could be reused as a reference taxonomy. For instance, one could take the Big Five trait definitions from an ontology and use their URIs as tags in our JSON schema. If a formal HEXACO ontology exists, that could likewise be adopted for the six-factor model. In practice, you might declare a simple JSON structure for a **"PersonalityProfile"** (with fields for each trait score) or embed trait scores as sub-objects. The key is that the trait names and definitions should align with an existing source (to maintain consistency). Even if we don't find a ready-made JSON schema for traits, we can map to known frameworks – e.g. using **IPIP** (International Personality Item Pool) facet names for trait categories or using well-known test constructs as identifiers. The IPIP itself provides a large repository of trait **facets** and questionnaire items (openly available), which can serve as an **open vocabulary** of personality attributes. For example, you might tag text with "IPIP:Friendliness" or "IPIP:ArtisticInterest" to indicate those facet tendencies. The structure could be as simple as an array of trait tags on each content item, or a more complex object linking to a trait ontology reference.

- **Narrative Identity and Life Story Metadata:** This area is more specialized, but you might consider ontologies that capture life events or narrative themes. While not a formal standard, concepts from

psychology (e.g. **“redemption narrative”**, **“attachment style”**, **“coping style”**) could be included as metadata on content. There isn’t, to our knowledge, an out-of-the-box JSON schema for “narrative identity.” However, you could extend an existing schema with custom fields or tags for these. For instance, add a field `narrativeTheme` to a journal entry object with values like “trauma”, “achievement”, or “family”. One could also use schema.org’s `about` property to link to WikiData items or a custom taxonomy of life events (for example, tagging a transcript with **about:** `SchoolGraduation` if it mentions that life event). If needed, simple ontologies for life events (some exist in sociology or semantic web contexts) can be leveraged. The W3C Provenance standard (PROV) or ActivityStreams “experience” types might model events in someone’s life log, but adaptation to narrative meaning would still require custom labeling. The **adaptation-layer metadata** like goals, values, coping styles are highly individual – a minimal way to include them is to have a profile object for the person with fields such as `coreValues: ["Independence", "Creativity"]` or `copingStyle: "Problem-focused"`. These could link to established psychological frameworks (e.g. Schwartz’s Values for values, or a known list of coping styles in health psychology), but likely as simple key-value pairs or tags.

- **Behavioral and Psychological Tag Vocabularies:** There are several lexicons that map language to psychological categories. **LIWC (Linguistic Inquiry and Word Count)** is a well-known proprietary example, where words are categorized (e.g. words indicating positive emotion, social processes, etc.). Many studies have used such categories to infer traits from text. For instance, higher *Neuroticism* correlates with using more negative emotion words and first-person singular pronouns, whereas *Extraversion* correlates with positive emotion and social words [12]. While LIWC itself isn’t open, there are open-source or research lexicons covering similar ground (e.g. the NRC Emotion Lexicon for emotions, or various academic datasets of personality-correlated words). An **open vocabulary** approach (as in the open-vocab personality studies) might use unsupervised topic extraction to find themes in one’s writings and then label those themes post-hoc. In a structured schema, one could incorporate these by having a field for **linguistic markers** or **inferred traits** on each content piece. For example: `"linguisticMetrics": { "negative_emotion_word_pct": 5%, "social_words_pct": 12% }` and perhaps `"inferredTraits": ["Neuroticism↑", "Extraversion↓"]` based on those markers. If a formal tagging vocabulary is desired, one could reuse **Schema.org’s** `DefinedTerm` construct or an existing psychological thesaurus. For instance, schema.org allows marking up keywords as either plain text or as a `DefinedTerm` with an identifier [9] – we could define a term set for personality facets. Similarly, the Big Five traits could be linked to Wikipedia/Wikidata entities (each trait like *Q9842 Openness to Experience* on Wikidata), and those URIs used as tags. This leverages open data so that any tool reading it knows it refers to that trait.

Recommendations for a Personality Cloning Pipeline

Considering the above, the most **suitable existing schemas** for this pipeline would likely be a hybrid of content-focused structure with an extension for personality metadata:

- **Use a Content Schema for Data Types:** Leverage a format like **ActivityStreams 2.0** (JSON-LD) or **schema.org’s CreativeWork** to represent the actual personal data items (entries, messages, transcripts). These are well-supported and can cover the variety of data (journal entry = an `Article` or `Note` object; email = an `EmailMessage`; chat log message = a `Note` or

`SocialMediaPosting`; meeting notes = `CreativeWork`; audio transcript could be a `CreativeWork` with a media type, etc.). ActivityStreams has the edge in being designed for social/personal activities and already has extension points (like the `tag` array and the ability to add custom fields in a context) [3]. This means we could attach personality-related info without breaking the structure. For instance, an ActivityStreams **Note** for a diary entry might include a custom property `"traits": {"openness": 0.8, "agreeableness": 0.6}` or a list of URIs in `"tag"`. Using schema.org alone is possible too, but you'd likely need to use its less structured keyword tagging and it's not as directly aimed at timeline events.

- **Incorporate Trait and Psych Metadata via Extension:** Whichever base schema above is used, we recommend a **lightweight extension** to encode personality and memory metadata. One approach is to define a JSON Schema or JSON-LD @context for a **"PersonalityAnnotation"** that can be attached to each item. For example, an extension could allow:

- tagging an item with trait scores or labels (Big Five trait scores 0–1, or discrete tags like "High Neuroticism" if qualitative),
- marking items as significant memories or linking them to user-defined goals/values.

If using ActivityStreams, you might add a property `"context"` or `"attachment"` that links to a PersonalityAnnotation object. In schema.org, you might use `mentions` or `about` to reference trait entities. The key is to reuse known vocabularies: for Big Five/HEXACO, use their trait names and definitions from psychology literature (ensuring consistency, e.g. always call it "Openness" and define as per Big Five). This could be as simple as a JSON file defining the trait names and maybe an ID for each (which serves as our mini ontology).

- **Minimal Extensions Needed:** Fortunately, none of the standards above require a completely new framework – they can be **adapted with minimal additions**. ActivityStreams already supports arbitrary extension fields via JSON-LD context [3], and schema.org can be extended with additionalType or custom sub-properties. We might only need to create:
- A list of trait terms (and possibly a mapping to numeric scores or categories).
- A way to indicate *why* a piece of text is tagged (optional: e.g. "this email was high in `{trait}` because it contains XYZ cues" could be stored if needed for interpretability, but not mandatory).

- Fields for "memory significance" or "user goal linkage" if those are part of the adaptation layer. For example, an extension like `"significance": "HIGH"` or `"relatedGoal": "Improve social skills"` can be added to entries to indicate these meta aspects.

- **Survey and Structured Data**: If your pipeline includes formal personality assessments (like the user taking a Big Five questionnaire), consider using **HL7 FHIR Questionnaire/QuestionnaireResponse** for that component [6]. This would allow encoding the questions and answers in a standardized way, and then you can convert the results into the trait profile. For general unstructured text, FHIR Observation is an option but might complicate things unless you are already in a health data context. If interoperability with health or research systems is a goal, you could map your ActivityStreams/schema.org data to FHIR profiles later. Otherwise, sticking to a JSON-LD content schema with custom tags is simpler for development.

In summary, we recommend **ActivityStreams 2.0** as a strong candidate for the core data format (to uniformly represent journals, messages, notes, etc.), augmented with a custom vocabulary for personality traits and related annotations. This choice is compatible with modern semantic web practices and can be easily extended to include Big Five/HEXACO tags. A similar alternative is using **schema.org CreativeWork** with `keywords`/`about` for traits (achieving essentially the same thing in microdata/JSON-LD format). Whichever base is chosen, only minimal extensions are needed: essentially, defining how to attach **trait scores/tags** and any **memory or goal metadata** to each item. By reusing open vocabularies (like IPIP trait names, or known psychological lexicons for emotions/values), we ensure the schema remains compatible with existing knowledge bases. This structured approach will enable a downstream large language model to ingest not just raw text, but rich metadata about the user's personality, thereby improving the fidelity of the digital personality clone.

Sources:

- W3C ActivityStreams 2.0 – JSON format for social activities [1].
- HL7 FHIR Observation – generic structured data point (incl. narrative/behavior) [4] [5].
- Open Humans platform – personal data files with tags/description metadata [7].
- Alamsyah et al. (2021) – ontology-based model for Big Five traits from social media [10].
- Kern et al. (2014) – linking linguistic markers (e.g. emotion words) to Big Five traits [12].
- Schema.org CreativeWork – supports tagging via keywords/about (semantic tags) [9].
- HL7 FHIR QuestionnaireResponse – represents answers to personality questionnaires [6].

---

[1] [2] Activity Streams 2.0. At long last, the Activity Streams 2.0... | by James M Snell | Medium
https://medium.com/@jasnell/activity-streams-2-0-70881f866935

[3] Activity Streams 2.0
https://www.w3.org/TR/activitystreams-core/

[4] [5] [6] Observation - FHIR v6.0.0-ballot2
https://build.fhir.org/observation.html

[7] Open Humans: A platform for participant-centered research and personal data exploration - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC6593360/

[8] Direct sharing - Open Humans
https://www.openhumans.org/direct-sharing/on-site-data-access/

[9] Schema.org CreativeWork Type tutorial - w3resource
https://www.w3resource.com/schema.org/CreativeWork.php

[10] Ontology Corpus for Big Five Personality Measurement - Andry Alamsyah Dataverse
https://dataverse.telkomuniversity.ac.id/dataset.xhtml?persistentId=doi:10.34820/FK2/L0FM7X

[11] Delivarable_D6.5
https://specialprivacy.ercim.eu/images/documents/SPECIAL_D65_M30_V10.pdf

[12] ASMNT-12-0254.R1 revision final 091313
https://www.peggykern.org/uploads/5/6/6/7/56678211/kern_et_al_2014_-_the_online_social_self_-_open_vocab_approach_to_personality_-_prepub_version.pdf