IOCM 2025 Conference

The 4th International Online Conference on Materials

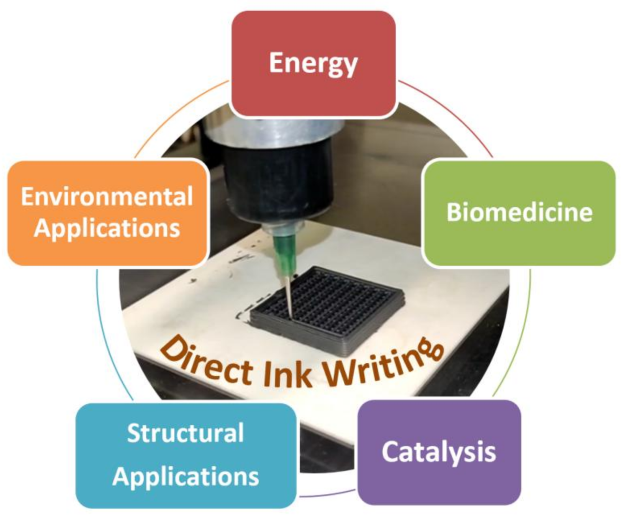
3–6 November 2025 | Online

MDPI

Development of a hybrid natural language processing system for the automated extraction of formulation data in direct ink writing

A. Bejerano , M. Belmonte and C. Ramírez
Instituto de Cerámica y Vidrio (ICV-CSIC), Kelsen 5, 28049 Madrid (Spain)

INTRODUCTION & AIM



- **Direct Ink Writing (DIW)** is one of the most versatile additive manufacturing techniques for developing cellular ceramic materials. It involves the extrusion of ceramic slurries (inks) with specific rheological properties to build 3D-structures layer by layer.
- ⚠ **Problem:** Designing printable ceramic inks for DIW heavily depends on trial-and-error, making the formulation process slow and labor-intensive.

- **Challenge:** Scientific literature addresses different technical aspects and is multimodal—mixing text, tables, figures, and sometimes code—making it difficult to extract the key parameters for standardization of formulations.
- **NLP** is a powerful tool for automatically extracting and structuring experimental information from scientific texts. Unlike simple keyword searches, it understands linguistic context, allowing more accurate identification of parameters and materials.
- **Aim:** To develop a ***hybrid NLP pipeline*** that can automatically extract formulation parameters from DIW literature, enabling: ***i)*** Systematic knowledge integration, ***ii)*** Reduced manual workload and ***iii)*** Optimized and ad hoc ink formulation design.

METHODOLOGY

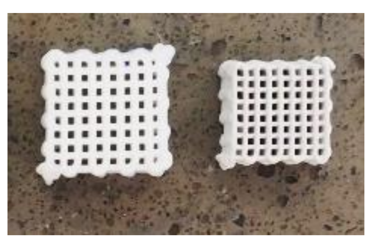
Key Parameters in Ink Formulation & Material Properties

Pre-Print Parameters

- Ceramic type (oxide, non-oxide...)
- Powder content (wt.%, vol.%, ratio)
- Powder morphology (spherical, fibre…)
- Particle size (μm, nm)
- Solvent type (organic, water…) and content
- Polymeric additives (dispersant, binder…) and content
- Rheological parameters (viscosity, elastic and viscous moduli, yield and flow points)

Printing Parameters

- Nozzle diameter
- Temperature
- Humidity
- Printing speed
- Extrusion pressure
- Design



Post-Print Parameters

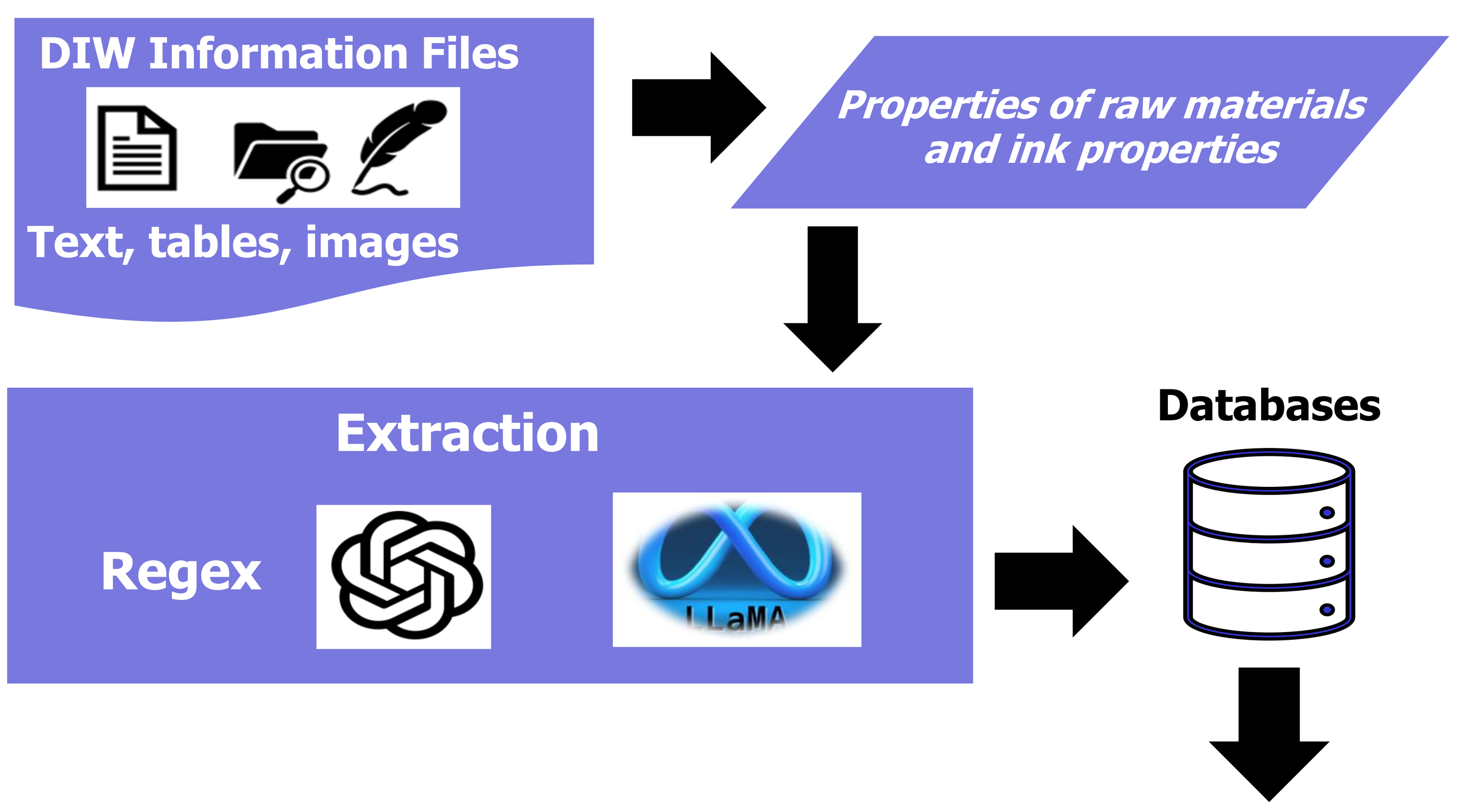
- Drying
- Debinding/calcination
- Sintering
- Structural properties (density, total porosity, filament porosity, strength)
- Functional properties

There is no fast, universal formulation for every material — that's why it's essential to **predict** and **optimize** these parameters.

PIPELINE CER3DML

A hybrid Natural Language Processing (NLP) system was implemented, combining:

1. Regular expressions for deterministic pattern matching.
2. Named Entity Recognition (NER) using language models to capture contextual entities.
3. Combination of pre-trained models such as GPT-4 or LLaMA for contextualization.
4. A manually curated subset of articles was used to validate entity recognition accuracy.



Optimization of printing parameters for ceramic ink **prediction**

RESULTS & DISCUSSION

Example of some variables in experimental procedure

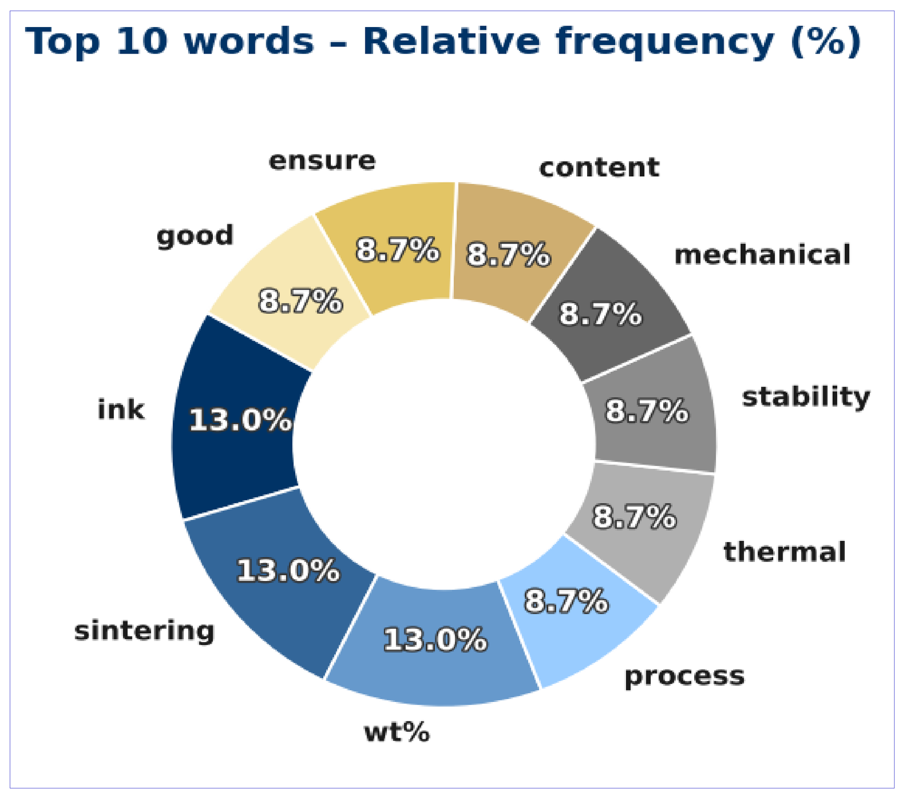
As a binder, Pluronic F-127 was incorporated at 25 wt%,
ble behavior allows efficient handling during the mixing
t extrusion. A 0.5 wt% of yttria was added as a sintering
ticle cohesion and promote uniform grain growth. The
(20 wt%) was carefully adjusted to achieve the desired

Pipeline Part I

🔍 **Text Extraction (PyMuPDF + PDFMiner + Camelot + Tabula)**

NLP Module (Regex + NER)

Example of recognition applied to a text with 22 variables



Pipeline Part II

NLP Module (GPT4 + LLAMA)

Field	Regex	LLAMA	OpenAI	Final	Source
Material	Alumina	Alumina	Alumina	Alumina	REGEX, LLAMA, GPT
Solid content(Vol%)	45 vol%	45 vol%	45 vol%	45 vol%	REGEX, LLAMA, GPT
Binder	binder	Alumina	Alumina	Alumina	LLAMA, GPT
Binder content (wt%)	25 wt%	25	25	25	REGEX, LLAMA, GPT
Additive	Alumina	PLURONIC	Yttria		REGEX, LLAMA, GPT
Water content (wt%)	25 wt%	25 wt%	25wt%	25wt%	REGEX, LLAMA, GPT

- Hybrid CER3DML pipeline: successfully extracted ~80% of key formulation entities and high robustness across heterogeneous sources.
- Regex modules: appropriate for deterministic values (e.g., percentages, viscosity)
- NER: captured contextual entities (e.g. polymer type. additive function), but exhibited semantic confusion between additives and base materials.
- Integrating context-aware reasoning models (GPT/LLaMA) should reduce these ambiguities.
- **Structured datasets enable quantitative comparison of ink compositions and can feed predictive models for printability and rheology.**

CONCLUSION

1. A hybrid NLP framework (Regex + NER + LLM reasoning) was developed for automated extraction of DIW formulation data.
2. The system bridges materials informatics and text mining, transforming unstructured scientific literature into machine-readable data.
3. Current limitations include entity overlap and semantic ambiguity, especially for additives and multi-component formulations.
4. Ongoing work focuses on semantic embeddings, transformers, and vectorized neural networks to improve context understanding.

ACKNOWLEDGEMENTS/ REFERENCES

- **Project MMT24-ICV-01** (funded by the European Commission – NextGenerationEU, through the Momentum CSIC Programme: "Develop Your Digital Talent")
- **References:**
- https://www.llama.com/
- https://openai.com/es-ES/api/pricing/