


<b>Cover Sheet for Proposals</b> <i>(All sections must be completed)</i>			
<b>Project Area:</b> <input checked="" type="checkbox"/> a) Interoperability Demonstrators			
<b>Name of Lead Institution:</b>		<b>EDINA National Data Centre</b>	
<b>Name of Proposed Project:</b>		<b>EM-LOADER</b> (Extracting Metadata to Load for Open Access Deposit)	
<b>Name(s) of Project Partner(s):</b>		<b>Textensor</b>	
<b>Full Contact Details for Primary Contact:</b>			
<b>Name:</b>	Peter Burnhill		
<b>Position:</b>	Director, EDINA		
<b>Email:</b>	p.burnhill@ed.ac.uk		
<b>Address:</b>	158 – 162 Causewayside, Edinburgh EH9 1PR		
<b>Tel:</b>	0131-650-3301		
<b>Fax:</b>	0131-650-3308		
<b>Length of Project:</b>	12 months		
<b>Project Start Date:</b>	March 2008	<b>Project End Date:</b>	February 2009
<b>Total Funding Requested from JISC:</b>		£29,882	
<b>Total Institutional Contributions:</b>		£30,947	
<b>Outline Project Description</b>			
<p>This project will demonstrate middleware that enables easier deposit of research papers through batch upload of extant bibliographic metadata. This will contribute to the work of the CRIG in the provision of shared infrastructure for digital repositories, taking forward into practice ideas mooted for a 'deposit engine'. It will also have immediate practical value as this middleware can be employed to assist deposit into the Depot as well as offer facility for repositories more generally, with potential to enhance metadata deposit through transfers and re-directs to institutional repositories (IRs).</p> <p>Using a web service approach and m2m interfaces such as Deposit API / SWORD, this middleware facility will show proof of concept at an early stage by connecting two existing services: the Depot, a UK repository for researchers who do not have other provision, and PublicationsList.org, a web site for researchers to build a web page listing their publications. The latter has existing functionality for batch import of bibliographic metadata for a (personal) publications list - from a variety of online sources such as PubMed, Web of Science, and for the same for personal databases, such as EndNote, Reference Manager, BibTeX etc. Having reviewed the simple single-item deposit workflow, this project will demonstrate the value of acceleration across the workflow, gaining leverage for the depositor, and for repository services, from batch ingest from databases containing structured bibliographic data.</p> <p>The overall aim is to help populate repositories, by making it simpler and more rewarding for researchers to submit their papers into repositories. The approach taken here is two-fold: to reward potential depositors with useful bibliographies of their own publications; from this to leverage benefit through batch upload followed by deposit of the full text.</p>			
<b>I have looked at the example FOI form at Appendix A and included an FOI form in the attached bid (Tick Box)</b>	<input checked="" type="checkbox"/> YES	<b>NO</b>	
<b>I have read the Circular and associated Terms and Conditions of Grant at Appendix B (Tick Box)</b>	<input checked="" type="checkbox"/> YES	<b>NO</b>	

## ***FOI Withheld Information Form***

We would like JISC to consider withholding the following sections or paragraphs from disclosure, should the contents of this proposal be requested under the Freedom of Information Act, or if we are successful in our bid for funding and our project proposal is made available on JISC's website.

We acknowledge that the FOI Withheld Information Form is of indicative value only and that JISC may nevertheless be obliged to disclose this information in accordance with the requirements of the Act. We acknowledge that the final decision on disclosure rests with JISC.

<b>Section / Paragraph No.</b>	<b>Relevant exemption from disclosure under FOI</b>	<b>Justification</b>
	Subject to Data Protection Act 1998	The information in our Curricula Vitae is personal data provided exclusively for the purpose of evaluating this bid. We withhold permission to use of that information for any other reason.

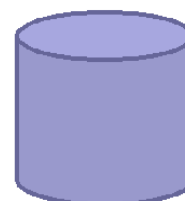
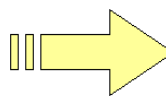
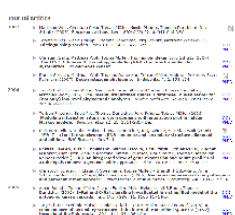
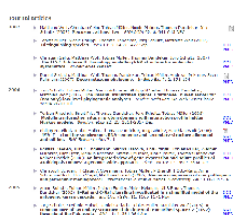
### Extracting Metadata to Load for Open Access Deposit: Integrating personal publications list management into repository submission

1. This proposal is submitted by EDINA as a provider of network-level repository services, including the Depot (part of JISC Repository Net), made with Textensor Limited, developers of PublicationsList.org. This project activity will contribute to the work of the CRIG in the provision of shared infrastructure for digital repositories, taking forward into practice ideas mooted for a 'deposit engine'. It has immediate practical value in assisting deposit into the Depot, with an approach that will enhance deposit generally, through use by repositories (IRs), interoperation with Current Research Information Systems (CRIS) and transfers to (and between) IRs.
2. The overall aim of this project activity is to help populate open access repositories, by making it simpler and more rewarding for researchers to submit their papers into repositories. One key part of the technical strategy is to place deposit of the digital object within a more established workflow of metadata creation, and in doing so to use that metadata to make deposit easier and more effective, both in terms of enhanced metadata and, by use of batch loading, acceleration across the single-item workflow. Another way of stating this is that the approach is two-fold: to give reward (and therefore added motivation) by assisting depositors compile useful bibliographies of their own publications; to leverage benefit from those publication lists through batch upload followed by deposit of the full text. The second key part of the technical strategy is to make use of existing functionality to enable researchers (potential depositors) to extract extant bibliographic metadata from a variety of online sources such as PubMed, Web of Science, and to do the same for personal databases, such as EndNote, Reference Manager, BibTeX etc. The third key part of the technical strategy is to show proof of concept, as a web service module that has m2m interfaces to support such as Deposit API / SWORD, at an early stage by connecting two existing services: the Depot, the JISC repository for researchers who do not have other provision, and PublicationsList.org, a web site for researchers to build a web page listing their publications. The Depot already makes API calls to such 'web services' for authentication, to OpenDOAR for its re-direct functionality, and to RoMEO; we envisage that the EM-LOADER would operate similarly utilising m2m/API access.
3. Depositing papers into repositories can be made easier and rewarding for researchers by concentrating initially on compiling a personal publications list (with complete metadata) and then performing a batch submission to the repository. The addition of the full text of papers (the digital objects) is possible at various stages in the process. Compiling a personal publications list prior to batch submission (rather than focussing immediately on submission of individual papers) is also useful to researchers in its own right. The step from here to a complete repository submission is clearly less onerous than direct submission of individual papers, as the metadata for papers is already present.
4. Typically Stage 1 in the figure above - compiling a personal bibliography – is by manual entry, but this can be

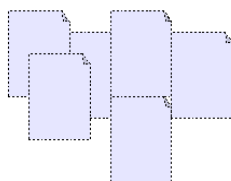
1. Researcher compiles list of their publications

2. Attaches full text where available (pdf / doc)

3. Batch submission to repositories



Batch metadata import from PubMed, Web of Science, bibtex, endnote, reference manager etc



made much easier with batch search and select of items from citation databases such as Web of Science and PubMed, and import from personal bibliography tools such as BibTeX, EndNote and Reference Manager. Clearly the end result of stage 1 is useful both for repository ingest and in its own right since most researchers require an up to date web-accessible list of their publications.

5. Full text of papers can be uploaded and attached to metadata in Stage 2 (typically in PDF or DOC formats). Note that it is preferable for researchers to upload the full text as part of this process. However, a researcher may not have access to the author's final preprint for their earlier papers, or it may take them some time to

find all the files, so it is likely that there will not be full text for all publications at this stage, but the completed metadata can be ready for when the digital object does become available.

6. Functionality for stages 1 and 2 already exists and is provided to this project through PublicationsList.org as a special feature of this proposal; the main focus of project activity is that part of the workflow in which this structured metadata are forwarded to the appropriate repository, with the associated digital object (full text) where available.
7. A small amount of additional metadata is required from the depositor before each paper can be made public. This includes checking journal copyright policies, embargo periods and open access policies, and in general does have to be completed as single-item workflow. These issues are already handled by the Depot's interface which makes calls to services such as SHERPA/RoMEO. Once the metadata and papers have been added to the Depot, they will be made available in the standard export formats supported by EPrints.

### Machine interfaces for submitting a publications list to a single repository

8. Once a publications list has been prepared, submitting all publications to a single repository will use a web service API such as Deposit API / SWORD, with appropriate authentication. We will do so initially in the context of submission to The Depot (EPrints), in order then to test this with the Edinburgh Research Archive (DSpace). This phase will require development of new modules for packaging up metadata and publications for use with Deposit API web service methods.
9. M2M interfaces are also required for transfer of digital objects and their metadata between institutional repositories. IRs typically only store the publications written by employees, so a researcher's publications may need to be routed to different repositories when employment changes. For this project we will allow selection of a subset of publications in a list followed by submission to a single repository. The Depot currently includes the capability to redirect users to their local UK repository, and the ability to transfer Depot contents onwards to other repositories is under development (independently of this proposal).
10. We expect to track OAI-ORE developments and make use of existing import modules / implementations of Deposit API / SWORD interfaces for eprints for performing the batch import. It is likely that some extensions to these interfaces will be required to keep the publications list and repository metadata synchronised when the user updates entries in either system. These extensions and other experiences with the integration process will be fully documented in the final report, as well as recording our experience of SWORD on the UKOLN wiki (<http://www.ukoln.ac.uk/repositories/digirep/index/SWORD>).

### Relation to the 'deposit scenario' described in the call for proposals

11. The JISC circular 5/07 described a hypothetical 'deposit engine' describing possible phases for the deposit process for a single paper. We have reflected upon the posited single-item workflow and identified both the requirement and the benefits of accelerating the deposit process through batch loading. Rather than a single paper, we start with a researcher's entire publications list, as might be found in a researcher's own or her group database/website, or even her CV if held in a structured format. Much of the technical work required is independent of the precise scenario, including: **a)** uploads of papers via the web; **b)** lookups in bibliographic databases and import from bibliography formats (already provided by PublicationsList.org); use of Deposit API / SWORD for batch submission; **c)** assertions on the author's right to make the paper available, based on the journal's open access policies, provided by The Depot's web interface.
12. The advantage of making an author's online publications list the starting point for deposit is that researchers are highly motivated to make their own web page and that of their research group look good: they perceive these as useful for marketing their research and typically the list will already be available as part of their C.V. Having assembled their publications list, the additional work required for deposit is limited to attaching pdfs and confirming their IPR rights to release the article through the OA repository. The Depot includes the additional checks on journal open access policies via RoMEO, which will help reassure researchers that most journals allow them to post eprints in open access repositories.
13. The value of this 'deposit scenario' in which a researcher first uses an online bibliography management tool is illustrated in two 'blogs'<sup>1</sup>. Active researchers have metadata details of their publications extant in some bibliography file (e.g. EndNote / Reference Manager) or by being able to fetch from a citation database such as Web of Science or PubMed. They should have the option to send all or part of the list to a particular repository. As there is motivation to keep such lists up to date, for new papers (or to attach the PDFs from earlier articles), we envisage a 'click button' to synchronise entries with the repository. This step will also update the publications list with links to the final repository entries.

---

<sup>1</sup> <http://dljtj.org/2007/06/two-personal-repository-services/>  
<http://david.davies.name/weblog/2007/07/08/publication-lists-and-eprints-self-archiving-with-publicationslist/>

## Plan of work and deliverables

14. The project will create, as a web service, the functionality available from the synergy of linking 'PublicationsList.org' with the Depot. This will involve use of the machine-machine interfaces developed by the Deposit API / SWORD group, and will require development effort at both ends to call the appropriate web service APIs and connect the two services. PublicationsList.org and the Depot are both operational services. We will implement the interconnection in a sandbox installation of PublicationsList and the Depot, and test the workflow on existing users of both services. We will engage with the CRIG throughout the project. Deliverables from the project will include reports on the experience gained from the technical aspects of using the APIs in earnest, as well as the user feedback on whether this approach to depositing papers in repositories is more attractive to researchers than existing web forms.

Qtr	Activity	Deliverables
Q1	Experiments and initial mashup using suitable APIs for batch submission of metadata from publicationslist.org to the Depot's installation of eprints.	<ul style="list-style-type: none"> <li>▪ Project website and collaboration infrastructure</li> <li>▪ JISC and eFramework project entries</li> <li>▪ Experimental mashup of publicationslist.org and the Depot for developers.</li> </ul>
Q2	Obtaining additional metadata from users on IPR / suitability for repository. Cross site authentication. Trials on users.	<ul style="list-style-type: none"> <li>▪ Working prototype of connection between two sites suitable for trials on end users.</li> <li>▪ Report from usability sessions.</li> <li>▪ Requirements document for additional features required from APIs for synchronising entries between bibliography management tool and repositories.</li> </ul>
Q3	Implementing synchronisation features. Plan for roll-out to other repositories. Working with CRIG support team to document this workflow.	<ul style="list-style-type: none"> <li>▪ Prototype of synchronisation features for keeping personal publications list and repository entries up to date with changes at either end.</li> <li>▪ Documentation of changes to APIs needed for use in this context.</li> </ul>
Q4	Reports and evaluation.	<ul style="list-style-type: none"> <li>▪ Short report that could be co-authored with CRIG support on lessons learned and additional requirements using APIs for synchronising personal publications list with repositories.</li> <li>▪ Final report.</li> </ul>

## Summary of benefits to the partners

15. This is opportunity for EDINA and Textensor to work together and engage in CRIG-related project activity. In addition, EDINA and the Depot project expect to benefit from this project by making the Depot facility substantially easier to use and demonstrate a new workflow which makes deposit more rewarding for researchers. This should lead to significantly more eprints being submitted via the Depot, and hence to extant institutional repositories, and thereby to other researchers. The main benefit of the project for Textensor Limited / PublicationsList.org is that this project activity raises its profile with repository staff and with JISC and the Open Access movement. It will make the PublicationsList.org site more immediately useful for institutional repositories, which may then be more likely to subscribe to the PublicationsList.org service. Because of this Textensor Limited will contribute £12,000 towards its total development costs of £33,000 for the project.

## Budget

16. This is a small-scale project accorded strategic value for obtaining experience in using deposit APIs in a real-world application and exploring the benefits for use by linking the two services on the Web. In order to reduce costs, EDINA has forgone claim on Indirect Costs, and EDINA and Textensor have both forgone claim on 30% of the directly incurred costs. EDINA will also underwrite any further necessary travel from other funds – typically, share of Indirect Costs for other related activity. There is also waiver, as the Directly Allocated Costs, of input from the Director of EDINA (as PI) and the EDINA Head of Bibliographic and Multimedia Services. { \*The institutional contribution includes all the directly allocated (D) and indirect costs (E), and 30% of the directly incurred costs (C) }



22. New users arrive at [publicationslist.org](http://publicationslist.org) via search engines and recommendations from co-authors. The PublicationsList.org site has been favourably reviewed in 'repository blogs' and has over 3,100 registered users worldwide, with UK users registered with \*.ac.uk email addresses including those from: Imperial College London, Universities of Bristol, Cambridge, Cardiff, Edinburgh, Essex, Leicester, Liverpool, Manchester, Nottingham, Oxford, Plymouth, Southampton and Warwick, Heriot Watt University, King's College London, Northumbria University, Open University, Queen Mary College University of London, The Royal Observatory Edinburgh [see also <http://publicationslist.org/organisations.html>].
23. Preparatory to writing this proposal an email was sent to UK academic users of PublicationsList.org to gauge their interest in using it as a front end for repository submission. This had responses such as:
- Q: Do you currently use your library's open access repository (if it exists) to make your papers available?* No. It seems to me such an open access repository doesn't exist. U. of Manchester
- Q: Would you be likely/willing to use a facility to submit your publications list to your library's repository?* Yes, certainly. With sincere thanks for your good service. U. of Manchester
- "My group recently started using PublicationsList.org to make our research available on the web and we're very happy with it. We don't have an institutional repository yet, but when we do, submission via publications list would be the ideal solution to getting papers into it." Heriot-Watt University
- "I think our group (certainly I would) be interested in using a repository such as this. I wasn't aware of [the Depot], and also I was unsure about our rights to post papers. The point about being able to post at least the authors' final copy is a good one."
24. We plan to approach these users to participate in the project by submitting their publications lists to/via the Depot once we have implemented the prototype batch submission system. Once we have refined the workflows, incorporating user feedback and suggestions, Depot submission will be offered as a standard feature for new users of PublicationsList.org. For users of the Depot, publications list management will be offered as an initial phase in preparing their submission.
25. We also wish to investigate how best to extend the results to extant UK institutional repositories, with help from CRIG members, either by depositing via the Depot and then making these items available to the appropriate institutional repository, or else by distributing the required EPrints / DSpace batch import / sync interface modules. Initially we plan to work with Edinburgh Research Archive [contacts: Dr Theo Andrew and Morag Watson], which apart from being local is based on DSpace repository software (the Depot is based on EPrints), so we will be able to have understanding of the two major repository systems used by UK institutional repositories.

#### **Interaction with developers of other services**

26. Accelerated deposition to repositories is of wide interest, including repository development teams (especially EPrints and DSpace), and the JISC CRIG, RRT and SWORD groups. We plan to align our initial APIs with the work done by these projects and with existing import modules (e.g. for BibTeX), but expect that there will be additional requirements which emerge from the need to synchronise a submission with an ongoing publications list. As we develop our prototypes we will publish the required APIs and modules to make them available to other projects.
27. It should be noted that the Depot already makes use of external web resources for authentication and checking journal preprint policies (including Athens, SHERPA/RoMEO, OpenDOAR). PublicationsList.org uses web services provided by PubMed (NIH), and has contributed bug fixes to the open source BibUtils framework for cross-converting bibliography formats.

#### **Technologies and open standards to be used**

28. The project is based on two sites: technologies used for PublicationsList.org include a front end in javascript / AJAX / DHTML / CSS and a server built using PHP and the Bibutils library for bibliography format conversions. It currently delivers structured data on publications to the browser in javascript object notation (JSON) format to make the user interface responsive and make best use of the browser cache to minimise server load. Bibutils offers conversion to / from bibliography formats including EndNote, BibTeX, Reference Manager, ISI, PubMed and MODS, and an Atom feed is also provided for news aggregators. We will investigate the most appropriate data transfer format to use for submitting bibliography details to the Depot; either JSON or an XML dialect (e.g. MODS) submitted using HTTP POST, possibly based around the Atom publishing protocol used by the SWORD group.
29. The Depot's repository interface is based on the open source EPrints code [based on Perl / MySQL], with a front end making use of javascript for a responsive user interface. For the workflow connecting the two, we will develop and document the web APIs needed to ensure that the result is a generic way of linking publications list management into the deposit process, not restricted to PublicationsList.org or the Depot.

30. **Evaluation / success of the project:** This will be measured by the user feedback of ease of use of the new workflows we develop, and the technical judgement of peers. We plan to offer new functionality to users within 12 months of project start.
31. **Risks:** The main technical risks are that the existing deposit APIs are not sufficient for the task of synchronising a publications list with repository entries, and that new interfaces will be required once we start connecting the two sites in practice. However, finding out new requirements for such APIs in real scenarios is part of the objective of this funding scheme, and the lessons learned will be relevant for the ongoing development of such APIs.
32. **IPR statement:** The project will have required access to the PublicationsList.org code (owned outright by Textensor Limited) and the Depot (based around the open source EPrints repository back end). All software, APIs and documentation funded by this project will be made open source with a LGPL or BSD style licence.
33. **Sustainability:** This is a small project for investigating the use of deposit APIs for making repository submission more effective. That said, this project will have practical and sustainable outcome in describing and enacting a workflow, to be assessed for inclusion in the Depot (which is supported by a long-term organisation, EDINA), and as an additional feature in PublicationsList.org site (which is supported by paid subscriptions). The workflow will also be applicable to interfacing to repositories from other online and desktop bibliography management systems.

## Project Team

### Project Management

34. Lead partner for the project is EDINA which will ensure that the project runs according to schedule and be responsible for developments connected with the Depot, including for project reporting to JISC. Textensor Limited will be responsible for developments connected with extensions and interfaces required for the PublicationsList.org site. We will have monthly face-to-face meetings to discuss progress, and will set up a password protected project website for internal discussion. We will use standard tools for version control (CVS / SVN) and for bug/issue tracking will use the Notate collaboration software developed by Textensor Limited. News items will be posted on a shared website.

### Previous Experience.

35. **EDINA** has established reputation with JISC for project competence as well as service orientation. This project is closely aligned with development plans at EDINA for the Depot and other JISC-sponsored repositories (such as Jorum and GRADE); it is part of its contribution to supporting scholarly communication: SUNCAT (national union serials catalogue) to UK OpenURL Router.
36. EDINA has proven project management and software engineering skills, and an appreciation of how to define and understand online communities of use. Led by Christine Rees (Head of Bibliographic Services) with 25 years computing service experience, the team includes Ian Stuart, the developer for the Depot, and Dr Theo Andrew, manager of Edinburgh Research Archive, who has recently joined EDINA as (half-time) project manager of the Depot, with EDINA's User Support Team to assist outreach and field-testing for the project. Peter Burnhill is director of EDINA (1995 -) with extensive experience in project leadership and oversight of projects that have been transformed into successful services. Having started on the science staff of SSRC (now Economic & Social Research Council), he was an active researcher from 1979 to 1987.
37. **Textensor Limited** brings to this project key skills that include an understanding of the requirements and motivations of the target users as well as an interest in developing and refining user interfaces using the appropriate mix of web technologies.
38. The PublicationsList.org site was launched by Textensor Limited in 2007 with the aim of making it as simple as possible for researchers to create and maintain a professional web page listing their publications. It makes unobtrusive use of javascript and AJAX techniques to deliver a simple but effective user interface which many of the 3,000+ users have responded to enthusiastically.
39. Fred Howell has a MEng in microelectronic systems engineering (UMIST) and PhD in computer science (Edinburgh University, 1996). After his PhD he worked as a postdoc researcher in computer simulation and then neuroinformatics, working on modelling tools and usable scientific database user interfaces for neuroscientists (one of which led to a spin-out company [www.axiope.com](http://www.axiope.com)). He left the university to form Textensor Limited ([www.textensor.com](http://www.textensor.com)) in 2005 with Robert Cannon, the company receiving a SMART R&D award from the Scottish Executive to develop new approaches to developing web based user interfaces for creating structured content from plain text ([www.textensor.com](http://www.textensor.com)). This has given him extensive experience of AJAX / Web 2.0 technologies for delivering a smooth user interface using a web browser.



11<sup>th</sup> January 2008

JISC Executive  
Northavon House  
Coldharbour Lane  
Bristol  
BS16 1QD

INFORMATION SERVICES

The University of Edinburgh  
Main Library Building  
George Square  
Edinburgh EH8 9LJ  
Telephone +44 (0)131 650 1000  
or direct dial +44 (0)131 650 4977  
Fax +44 (0)131 650 4978  
Email [jeff.haywood@ed.ac.uk](mailto:jeff.haywood@ed.ac.uk)

Dear Sir/Madam,

**JISC Circular 05/07: Call for Projects**

I am writing in support of the participation of the University of Edinburgh through EDINA in a bid being submitted in response to the above invitation to tender.

EDINA will lead the project which also includes Textensor Ltd, the developers of the PublicationsList.org. This bid combines the strengths and significant experience of the two partners in an exciting and challenging area.

Yours sincerely

Jeff Haywood

Vice Principal of Knowledge Management, Chief Information Officer, and Librarian  
to the University

JISC  
Northavon House  
Coldharbour Lane  
Bristol, BS16 1QD

Textensor Limited  
37 McDonald Road  
Edinburgh  
EH7 4LY  
UK

[fred@textensor.com](mailto:fred@textensor.com)

tel: 0131 332 3846

12th January 2008

**Letter in support of PublicationsList.org / EDINA Depot JISC Grant Application**

Textensor Limited, the developers of the PublicationsList.org site, are very enthusiastic about this collaboration with EDINA's Depot project to investigate a new workflow using PublicationsList.org for easier batch deposit of research articles into repositories.

The PublicationsList.org site has proven very popular with researchers around the world since its launch in April 2007 (over 3000 researchers now use it, with metadata and full text links for over 30,000 publications). It provides a web based user interface which makes it simpler for researchers to keep their personal and group publications list online and professional. Developing workflows which connect PublicationsList.org features with institutional repositories provides an excellent way for UK PublicationsList.org users to enhance their publications list with links to full text open access copies in their institutional repository. The Depot project has developed essential and complementary features for supporting deposit into institutional repositories, including Athens authentication, OAI-PMH interfaces, and checks with SHERPA/RoMEO database.

We expect that the outcome of the project will be to make the process of adding publications to an institutional repository simpler and more rewarding for researchers, resulting in substantially more research material being available open access and we look forward to collaborating with EDINA to make it happen.

Yours sincerely,

A handwritten signature in black ink that reads "Fred Howell".

Dr Fred Howell

Director,  
Textensor Limited